CalligraphicOCR for Chinese Calligraphy Recognition

Xiaoyi Bao[1], **Zhongqing Wang**[2*], **Jinghang Gu**[1*], **and Chu-Ren Huang**[1]
[1]The Hong Kong Polytechnic University, Hong Kong, China
[2]Natural Language Processing Lab, Soochow University, Suzhou, China
p2213545413@outlook.com
wangzq@suda.edu.cn
{jinghang.gu,churen.huang}@polyu.edu.hk

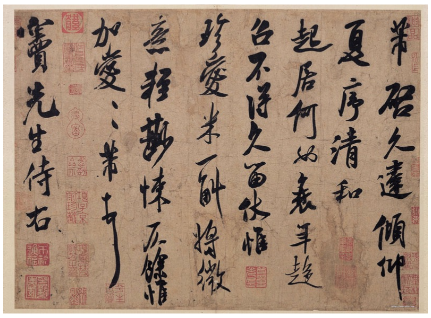
Abstract

With thousand years of history, calligraphy serve as one of the representative symbols of Chinese culture. Increasing works try to digitize calligraphy by recognizing the context of calligraphy for better preservation and propagation. However, previous works stick to isolated single character recognition, not only requires unpractical manual splitting into characters, but also abandon the enriched context information that could be supplementary. To this end, we construct the pioneering end-to-end calligraphy recognition benchmark dataset, this dataset is challenging due to both the visual variations such as different writing styles, and the textual understanding such as the domain shift in semantics. We further propose CalligraphicOCR (COCR) equipped with calligraphic image augmentation and action-based corrector targeted at the challenging root of this setting. 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. The code and data are available at https://github.com/HoraceXIaoyiBao/COCR-EMNLP2025

1 Introduction

The history of Chinese calligraphy is extensive, from its earliest carrier on silk, bamboo, and textile scrolls to later works on paper and stone steles, calligraphers have created numerous works in diverse writing styles, among which exist heirloom classic works such as *Lantingji Xu* (兰亭集序) and *Eulogy for My Nephew* (祭侄文稿). These calligraphic masterpieces hold profound significance in shaping Chinese people cultural identity (Wang et al., 2020; Bao et al., 2025) and nature (Su et al., 2022).

米芾《清和贴》
北宋，公元1103年
Mi Fu, On Pleasant Harmony, Song Dynasty, 1103 A.D.



Input: Calligraphy Image

芾启。久违倾仰，夏序清和，起居何如？衰年趋召，不得久留，伏惟珍爱。米一斛，将微意，轻鲜悚仄。馀惟加爱、加爱。芾顿首。窦先生侍右。

(Long absent, I miss you deeply. Summer is serene, how fare you? Summoned by duty in old age, I cannot stay. A humble gift of rice conveys my regard. Take care.)

Output: Recognized Context

Figure 1: Illustration of calligraphy recognition.

However, while many people enjoy and practice calligraphy, very few have digitized it, putting it in a low-resource situation. Previous efforts include recognitions rely on CNN architecture (Huang et al., 2022), transformers (Dan et al., 2022) or the unique characteristics of Chinese (Chen et al., 2021; Bao et al., 2024). Despite their effectiveness, previous works' modelings are inapplicable to calligraphy because of their sticking to isolated character recognition (Carlson et al., 2024), requiring expensive manual splitting the calligraphy into single characters (Liu et al., 2013; Peng et al., 2022), also discard the contextual semantic information that is no less important than the visual shapes.

In this paper, we propose a new task: end-to-end calligraphy recognition, as shown in Figure 1, the input of our new setting is the complete calligraphy image and the output is the contexts in the calligraphy. Our task aims to guide practi-

*Zhongqing Wang and Jinghang Gu are the corresponding authors

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4865–4877
November 4-9, 2025

cal modeling methods for digitizing calligraphy works, thereby furthering the preservation of ancient calligraphy and supporting the construction of traditional Chinese cultural symbols.

To effectively benchmarking this task, we construct a dataset named Chinese Calligraphy Recognition (CCR). On the basis of classic calligraphy images, we build the dataset by hiring naive speakers to annotate the sentences in the image, which include the calligraphy works written by 91 calligraphers, with a time span of 10 dynasties from Wei (魏) to Ming (明), across 1,500 years. Our annotation is designed to cover corner cases as many as possible, the perspectives include the variations of different writing styles from neat (i.e., Slim Jin 瘦金体) to scribble (i.e., Huang 黄庭坚), the topics from government documents to love letters, formations from poems to diaries, and even with the stamps that could disturb the recognition. Thereby our CCR can facilitate the exhaustive benchmark of calligraphy recognition task.

However, it is challenging to recognize calligraphy image. As shown in Figure 2, the first challenge arises from the visual modality, encompassing: 1) **Diverse writing styles** stemming from individual habits, such as the Slim Jin (瘦金体) is famous for its neatness but Huang (黄庭坚) is scribble, having characters naturally joined-up and overlapped. 2) **Absence of segmentation** in calligraphy, leading to the missing of punctuations and random line breaks, having characters being written in an unsegmented, continuous manner. 3) **Noise artifacts** which could include seals (印章) and inscription (落款) that could seriously disturb the recognition. Additionally, the shift from isolated character to complete context introduces the second challenge: how to utilizing the textual context semantics to reinforce the recognition under the serious 4) **Domain shift**, where the language expression in calligraphy may changes over thousand years of evolution while current language models are not familiar with it.

In this study, we address the above challenges with the proposed CalligraphicOCR (COCR). As shown in Figure 4, our approach combines calligraphic image augmentation and an action-based corrector. The former gradually refines the training images to closely resemble real calligraphy works through three augmentation strategies, while the latter contains a concise set of editing actions simulate the human correction process and a novel alignment method to maximize the effectiveness of corrections, finally revised the output sentence with contextual semantics and distinguish our model from previous pure-visual recognitions.

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

2 Related Work

2.1 Chinese Calligraphy Recognition

Optical Character Recognition (OCR) aims to convert text in images into an editable format. OCR models can generally be categorized into two approaches: Traditional OCR (Liao et al., 2022, 2016; Liu et al., 2019), which is composed of multiple expert modules, and VLM-driven OCR (Bai et al., 2023; Liu et al., 2024; Chen et al., 2024b), whose capabilities are derived from CLIP-style modules (Radford et al., 2021). However, the majority of them are focused on scene text or document recognition, the sparse works on calligraphy somehow trend to focus on single character recognition (Liu et al., 2013; Peng et al., 2022; Xu et al., 2019), such isolated recognitions are unpractical as they require unaffordable labor cost of manually splitting the calligraphy into characters, not to mention reforming them into readable sentences.

Unlike previous works, our benchmark and model stand out as the first to focus on the practical setting of end-to-end calligraphy recognition, thereby guiding the holistic optimization on this real-world challenges.

2.2 Post-correction for OCR

Post-correction for OCR has been extensively studied in high-resource languages, started from lexical techniques and weighted finite-state models (Schulz and Kuhn, 2017) to generations: Rijhwani et al. (2020) use a BiLSTM for historical English text, and Dong and Smith (2018) propose a multi-source model combining first-pass OCR outputs from duplicate English documents.

In contrast, research on lower-resource languages is limited. Anastasopoulos and Chiang (2018) leverage high-resource translations to improve low-resource speech transcription. Krishna et al. (2018) demonstrate OCR improvements for

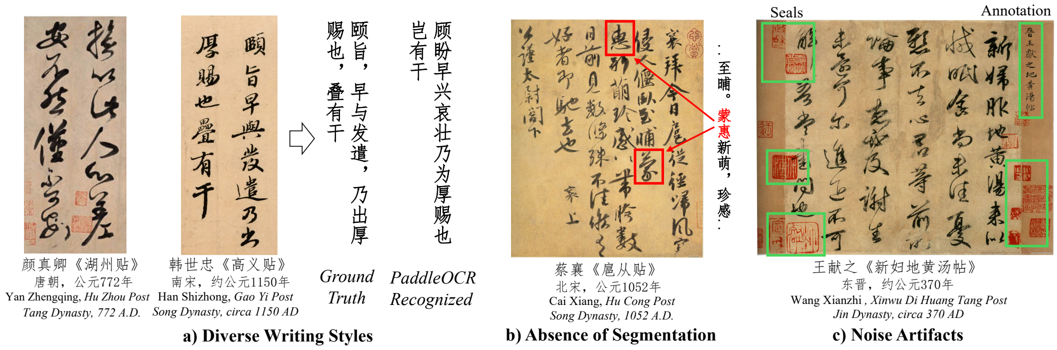


Figure 2: Illustration of visual challenges.

	Train	Dev	Test
#Samples	2500	227	200
#Chars / Samples	86.18	83.92	102.01
#Punctuations / Samples	14.30	14.77	15.65
#Authors	469	58	91
#Samples / Authors	5.33	3.91	2.19
#Dynasties	13	9	10
#Authors / Dynasties	36.07	6.44	9.10

Table 1: Statistics of our CCR dataset.

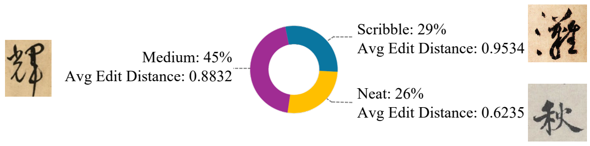


Figure 3: Statistics of neatness levels in our testset.

Romanized Sanskrit; Rijhwani et al. (2020) focus on endangered languages Yakkha and Nepali; Drobac et al. (2017) focus on Finnish.

Despite their effectiveness, our work distinguishes itself by being the first to concentrate on Chinese calligraphy recognition, thereby building a solid foundation for the downstream study on this low-resource language.

3 Task and Dataset

3.1 End-to-End Calligraphy Recognition

As shown in Figure 1, we first define the input is the complete calligraphy image solely without any text. The output will be the context in the calligraphy, seals, inscriptions and notes are not included in our target. The output should be readable and segmented sentence. Our task is then formulated as extracting a sequence of text elements

$$T = [t_1, t_2, \ldots, t_m] \quad (1)$$

from an input image I, where each t is a character identified in the image.

3.2 Dataset Collection and Annotation

We construct a new dataset called **C**hinese **C**alligraphy **R**ecognition **(CCR)** for benchmarking. CCR focuses on one of the most challenging and practical cases of end-to-end calligraphy recognition, thereby facilitating the solid benchmarking for downstream evaluation.

For the train and dev set, we collect the context of 2,727 classic Chinese literary works and sample 2,500 for the train set, the remained 227 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 200 calligraphy. Specifically, we hired native speakers to collect calligraphy samples from *calligraphy space* (书法空间)[1]. To simulate practical scenarios, we ensure annotation quality by applying the following standards: 1) Each sample must contain a minimum of 20 and a maximum of 200 characters. 2) Only complete, single-image calligraphy pieces are accepted; partial or cropped images are excluded. 3) Only calligraphy works from the dynasties spanning from Wei (魏) to Ming (明) are included. Works from earlier or later periods are excluded due to being either too ancient or modern.

3.3 Dataset Statistic and Analysis

We show detailed statistics of our CCR data in Table 1. We can tell that there are average around 15 punctuations per sample, which are missed in the calligraphy image and post a hard challenge

[1] http://www.9610.com/index1.htm

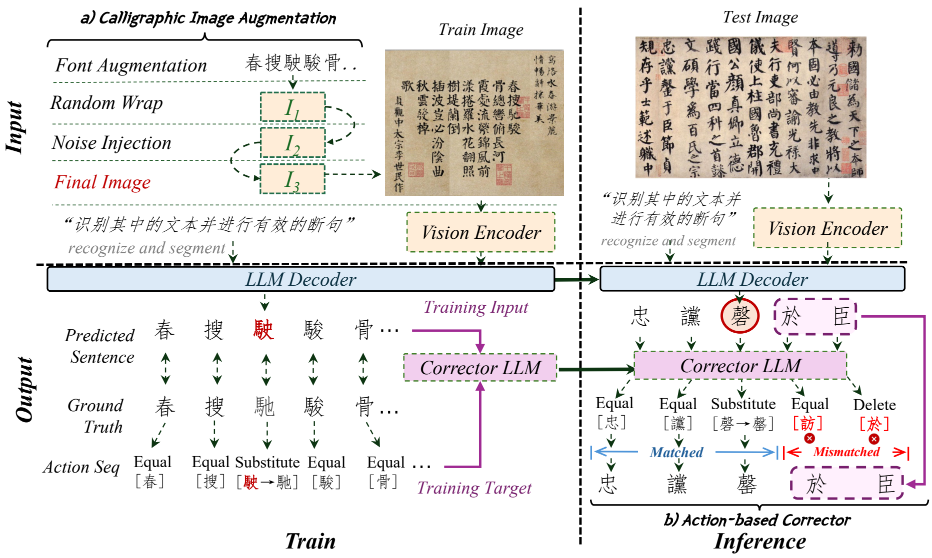


Figure 4: The illustration of our COCR.

for the recognition system to recover the punctuations properly. Besides, we also ensure the diversity of writing styles by extending the author and dynasties pool as large as possible, especially in the testset where only around 2 works per author.

To quantitatively measure the difficulty of recognition, we further divide our testset into three levels of neatness based on the average minimum edit distance per character between PaddleOCR (Du et al., 2021) output and ground truth. The calligraphy ≥ 0.9 are classified as scribbled, 0.9 to 0.7 are the medium, and ≤ 0.7 are the neat. As shown in Figure 3, 58 samples are classified into scribbled as the hardest level for recognition, 89 into medium, and 53 into neat. We will analyze the performance across three different levels in the experiment section.

3.4 Challenges

The challenges of our task lay in the two modalities towards the recognition, the first is the ***visual variations***, includes three aspects:

- **Diverse writing styles**: As shown in Figure 2 a), the right style features neat, clear typography, while the left is wild, with varying word sizes and overlapping characters. We show the widely used PaddleOCR (Du et al., 2021) struggles with this task.
- **Absence of Segmentation**: This leads to two key recognition challenges: the absence of punctuation and random line breaks. As illustrated in Figure 2 b), the second column shows an confusing blank space between 蒙 and 恵 that belong to the same clause.
- **Noise artifacts**: Besides from being blurred due to poor storage, there are noises are added deliberately by the depositories or authors such as the name seals (姓名章) and annotation shown in Figure 2 c).

The second challenge arises due to the shift from isolated character to complete calligraphy recognition, where the contextual semantics become available. The challenging point here can be summarized as:

- **Domain Shift**: The modern and classic Chinese could have significant difference on the expressions, for instance:

  "予除右丞相兼枢密使"
  (I was appointed as the Right Chancellor and Minister of the Imperial Secretariat.)

  the meaning of the word "除" (appoint) is different from its meaning of *remove* in modern Chinese. Such a domain shift could hinder the application of contextual semantics.

4 CalligraphicOCR

4.1 Basic Workflow

In this study, we propose a novel CalligraphicOCR (COCR). As shown in Figure 4, we follow the typical workflow of large vision-language models:

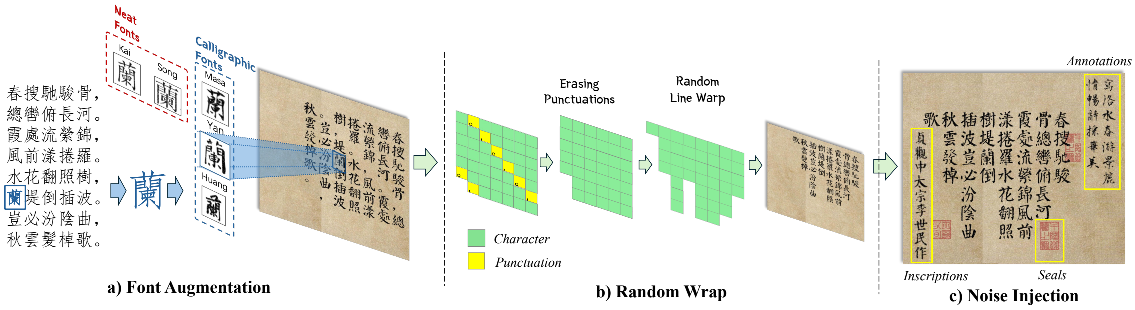


Figure 5: The illustration of our Calligraphic Image Augmentation.

when provided with calligraphy image and instruction, the LLM processes the vision encoder's output and concated it with the text as the input, the output target would be segmented sentence recognized. We then address the challenges by introduced two key components: Calligraphic Image Augmentation works on the input end, followed by Action-based Corrector at the output end.

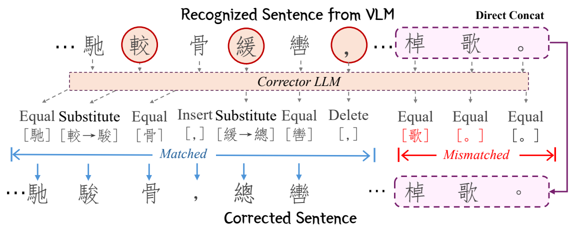


Figure 6: Our action-based corrector.

4.2 Calligraphic Image Augmentation

As shown in Figure 5, we propose three strategies to augment the input image in the train set in a pipeline manner to come close to the real calligraphic image step by step, each of them corresponds to one aspect of visual variations.

Font Augmentation

We first deal with different writing styles. Current pretrained vision-language models unable to cope the various writing styles because it has only been exposed to the standard fonts such as Song(宋体), which although covers the clear character structure but significantly lack the generalization towards scribble characters. We thus propose Font Augmentation method using two font sets. As shown in Figure 5 a), the first set consists of neat fonts, like Song (宋体), representing standard characters. The second set includes calligraphic fonts, such as Huang (黄庭坚), capturing writing styles beyond neat fonts. Each training image is re-rendered with one font from each set to improve the model's generalization to varied writing styles.

Random Wrap

We subsequently address the absence of segmentation, which leads to two difficulties: The missing of punctuations and random line breaks. We thus mock these writing habits by our random wrap to make sure as close as possible to the test images.

Specifically, we render training images by removing all punctuation while preserving the original word order and applying random line warping, where the next character is randomly placed either on the same line or on a new line to the left. This ensures that line breaks in the image do not indicate real segmentation and requires the model applying semantically compliant segmentation.

Noise Injection

We then 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 calligraphy are considered as noise such as seals and annotations. 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, include 1) Seals inserted with Seal Script (篆体) that randomly appear at any position in the image; 2) Annotations are generated by LLM with a prompt instructed to introduce the calligraphy, with a deliberate different font to distinguish from the author's scripts; 3) Inscriptions are generated similar to previous one, but printed in the same font.

4.3 Action-based Corrector

Shifting from single-character to full-text recognition brings the bonus of contextual semantics, which are often wasted in purely visual models, they can misrecognize characters that are totally incoherent with the context. We thus are motivated to explore a new way that can correct these errors with contextual information. We design an Action-based Corrector with a set of edit actions that emulate the how human editor act with the errors in the sentence. We then finetuned a generative LLM to generate the Correct Action Sequence on the basis of the recognized sentence and finally apply the actions on the sentence with our Action Alignment.

Correct Actions

As shown in Figure 6, we follow the edit action in Levenshtein Distance, design four edit actions to process the recognized sentence character by character, matching each character with an edit action, specifically include:

Insertion(A) // Insert char A
Deletion(A) // Delete char A
Substitution(A,B) // Replace char A with B
Equal(A) // Accept char A

where the A and B indicating the parameter of the action. We then fine-tune an LLM to generate action sequences for recognized sentences, using the VLM's output with possible errors as input. The output is organized based on minimal edit actions between the recognized sentence and the correct label, calculated in a dynamic programming approach of Levenshtein Distance.

Edit Action Alignment

After generating the action sequence, we need an effective method to align the actions with corresponding characters, any mismatching will propagate and accumulate offsets in the following alignments, making the edited sentence unreadable.

As shown in Figure 6, we use an alignment method to maximize valid actions. The action sequence $A = [a_1, a_2, \ldots, a_n]$ is matched to the text sequence $T = [c_1, c_2, \ldots, c_m]$ as follows: For pairs of $\{[a_0, c_0], [a_1, c_2], ...\}$ matched from A and T, the algorithm applies a_i to c_j one by one to iteratively update the corrected text T'. This continues until $i > n$, $j > m$, or an invalid action occurs. The corrected text T' is updated as:

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

where a_i will be judged valid with c_j and $\zeta = \{\text{Equal}, \text{Deletion}, \text{Substitution}\}$ by:

$$\text{f}(\text{a}_i, c_j) = \begin{cases} \text{Valid}, & \text{if } (a_i \in \zeta) \wedge (a_i.p = c_j) \\ \text{Valid}, & \text{if } (a_i \notin \zeta) \\ \text{Invalid}, & \text{otherwise} \end{cases} \tag{3}$$

where $a_i.p$ represent the parameter of action a_i, Upon an invalid action at position j, T' is formed by concatenating the corrected characters up to $j-1$ with the uncorrected characters from T:

$$T' \leftarrow T'[1 : j-1] + T[j : m] \tag{4}$$

This alignment ensures T' is constructed by maximizing valid actions while handling mismatches.

5 Experiment

5.1 Dataset and Experiment Setting

We evaluate the performance of our COCR and other baselines systems on the proposed datasets. For our Vision-Language Model in our COCR, we employ the pre-trained InternVL2.5-8B (Chen et al., 2024a) and LoRA fine-tune the LLM adapter parameters for 30 epochs. We adopt a LoRA finetuned Qwen2.5-7B for our corrector. All the Chinese characters in both the training images and texts are in traditional formation. Experiments were performed with four Nvidia A6000s.

We adopt commonly used metrics in OCR tasks, include F1-score, Character Error Rate (CER), and BLEU as previous works did (Wei et al., 2024a; Yousef and Bishop, 2020). Among them, F1-score is calculated over the recognized characters, focusing only on each character's recognition, not on sentence; CER is calculated by the average minimum edit distance per character and, together with BLUE, measures both single-character and sentence order recognition.

5.2 Main Result

In Table 2, we present a comprehensive comparison with cutting edge baselines, include: traditional OCRs: 1) PaddleOCR (Du et al., 2021), 2) EasyOCR (JaidedAI), 3) EffOCR (Carlson et al., 2024), VLM-driven OCRs, include off-the-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) InternVL2.5-7B (Chen et al., 2024a); 6) LLaVA-1.5-7B (Liu et al., 2023). Besides, we also

Method	↑ P.	↑ R.	↑ F1.	↓ CER	↑ BLEU
Human Baseline					
Human	0.6642	0.5393	0.5952	0.6218	0.1160
Traditional OCR Baselines					
PaddleOCR(off-the-shelf) (Du et al., 2021)	0.4579	0.3369	0.3740	0.9133	0.0035
EasyOCR(off-the-shelf) (JaidedAI)	0.4421	0.3016	0.3585	0.9218	0.0023
EffOCR (Carlson et al., 2024)	0.4072	0.4346	0.4204	0.8738	0.0513
VLM-driven OCR Baselines					
GPT-4o(off-the-shelf) (OpenAI, 2024)	0.6432	0.5410	0.5748	0.6718	0.0948
Deepseek-VL2(off-the-shelf) (Wu et al., 2024)	0.6175	0.5628	0.5888	0.6528	0.1031
GOT-OCR2.0 (Wei et al., 2024b)	0.4011	0.2111	0.2766	0.8767	0.0012
Vary (Wei et al., 2023)	0.4124	0.2466	0.3086	0.8918	0.0004
LLaVA-1.5-7B (Liu et al., 2023)	0.0113	0.0043	0.0063	0.9970	0.0001
Qwen2-VL-7B (Wang et al., 2024)	0.5134	0.5423	0.5274	0.6410	0.0422
Qwen2.5-VL-7B (Yang et al., 2024)	0.5323	0.5496	0.5408	0.6229	0.0537
InternLM-XComposer (Dong et al., 2024)	0.4967	0.5478	0.5210	0.7914	0.0337
InternVL2.5-8B (Chen et al., 2024a)	0.6959	0.5549	0.6174	0.6179	0.1076
Ours	**0.7037**	**0.6421**	**0.6715**	**0.5318**	**0.1326**

Table 2: Comparison with baselines.

Method	↑ F1.	↓ CER
Basic	0.6174	0.6179
Calligraphic Image Augmentation		
+Font Augmentation	0.6384	0.5746
+Random Wrap	0.6226	0.6034
+Noise Injection	0.6179	0.6092
+*All*	**0.6520**	**0.5549**
Action-based Corrector		
+Correct Actions	0.6278	0.6037
+Correct Actions, Actions Alignment	**0.6209**	**0.5939**
Ours	**0.6715**	**0.5318**

Table 3: The result of ablation study.

have human-recognized result by hiring 10 native speakers to manually recognize the testset, tasked 20 samples each person.

From Table 2 we can tell that all of the baselines show a noticeable low performance, indicating the difficulty of our task. Among the baselines, the VLM-driven OCRs such as InternVL2.5 outperform previous traditional OCRs, achieving a level close to human, these results highlight the effectiveness of the unified generation architecture, which can utilize the rich label semantics by encoding the natural language label into the output.

Furthermore, our proposed model demonstrates substantial improvements over all previous studies ($p < 0.05$). This underscores the effectiveness of our COCR framework when applied to calligraphic images. Particularly our model further surpasses the human result with a noticeable gap, validating our motivation to address the inherent challenges through the integration of augmentation and correction. We further show our model is generalized to standard fonts in Appendix A.

5.3 Ablation Study

We then investigate the contribution of our calligraphic image augmentation and action-based corrector. We use "Basic" to refer to the removing of two components, relying solely on the raw image.

As depicted in Table 3, when using only raw images, the performance is notably low, which is excepted since the VLM is not pre-trained on calligraphy image. Significantly improved performance is observed when the calligraphic image augmentation is included, we attribute this as it reinforces the robustness and generalization towards the calligraphic images. Furthermore, our action-based corrector, which, instead of sticking to pure-visual solution, aggregates context semantics into recognition and redeems the semantically outrageous errors, further enhancing the performance.

6 Analysis and Discussion

6.1 Impact of Fonts

We further investigate which type of font in our font augmentation can benefit the recognition more. Particularly, we train our COCR with two sets of fonts for the trainset: 1) Neat fonts that are more considered to be formal and standardized such as Song (宋). 2) Scribbled fonts are close to the calligraphy such as Huang (黄庭坚).

As shown in Table 4, performances within each group are similar. Between the two groups, neat fonts significantly outperform scribbled ones. This aligns with real-world teaching practices, where standard fonts are preferred for their clarity to convey character structure, enhancing gen-

Font	Type	Illustration	↑ F1.	↓ CER
Song(宋体)	Neat	蘭亭集序	0.6522	0.5891
Kai(楷体)		蘭亭集序	0.6539	0.5857
Mi(米芾)	Scribbled	蘭亭集序	0.6421	0.5956
Huang(黄庭坚)		蘭亭集序	0.6341	0.6015
Masa(正风)		蘭亭集序	0.6451	0.5997
Song(宋体) + Mi(米芾)	Mixed	蘭亭|集序	0.6638	**0.5427**
Song(宋体) + Huang(黄庭坚)		蘭亭|集序	**0.6672**	0.5492

Table 4: Result of different fonts.

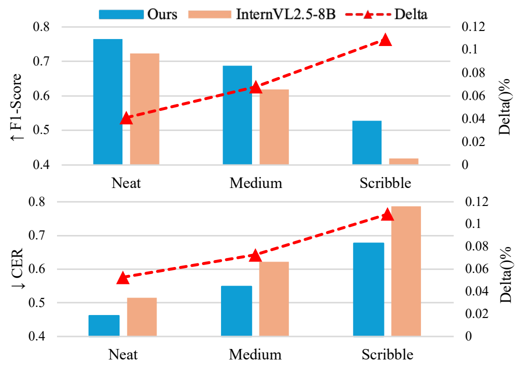


Figure 7: Results for different neatness levels. The top one is measured by F1-score, higher is better, while the bottom one is measured by CER, lower is better.

eralization to varied styles. In contrast, scribbled fonts like Huang (黄庭坚) mainly help recognize a specific style with limited generalization. However, combining Song (宋体) with scribbled fonts further improves performance, supporting our hypothesis that scribbled fonts complement corner cases under the broad generalization of neat fonts.

6.2 Impact of Calligraphy Neatness

We further investigate our proposed COCR's effects in different levels of neatness annotated. Specifically, we compare our method's performance with the strongest baseline across the three neatness levels in Figure 7.

We find that the more scribbled the calligraphy is, the lower the performance, which is expected since the scribbles in calligraphy pose obstacles to recognition and hinder the final performance. Moreover, the more scribbled the calligraphy is, the larger advantage our model has, we attribute this to our font augmentation, which brings our COCR the superiority in the difficult cases.

Additionally, we also analyze the impact of character formations in Appendix B.

7 Case Study

We launch case studies to make a more intuitive comparison between our COCR and the strongest baseline InternVL2.5-8B in Table 5.

We show that our COCR can effectively handle the scribble recognition cases in the first example, where the baseline encounter tough situation, outputs mojibakes while our COCR successfully recognize the target. In the second example, we illustrate that our COCR also performs better in the neat cases: the baseline get errors in both the segmentation and characters, whereas our COCR successfully avoids the problems above and helps the final recognition. We add more cases in Appendix C for more comprehensive illustration.

8 Conclusion

In this study, we highlight previous calligraphy recognitions are inapplicable to real-world situation and thereby hinder the preservation of Chinese calligraphy. We thus propose a novel task: end-to-end calligraphy recognition that aims to recognize readable segmented sentence from classic Chinese calligraphy work at one stop. We further propose Chinese Calligraphy Recognition dataset to fulfill the evaluation. With our calligraphic image augmentation and corrector, our COCR builds a strong benchmark for our task and effectively promote the preservation and dissemination of calligraphy.

Limitation

The limitations of our work can be stated from two perspectives. Firstly, the source of calligraphy works are limited, more sources such as bamboo slips(竹简), frottages(拓印), stele inscriptions(碑文) and oracles are still unexplored. Further exploration on more possible sources, especially combined with historic background could provide valuable insights.

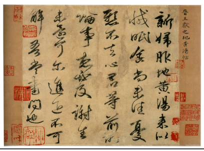
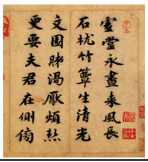
Input Calligraphy Image	InternVL2.5-8B	Ours	Ground Truth
	新婦服地黄湯來以🄵事。 壹或反謝生未還🄵進退。 不可解吾常未問迪。	新婦服地黄湯來，似減。 眠食尚未佳，憂懸 不去心。君等前所論事， 想必及。謝生未還，可爾。 進退不可解，吾當書問也。	新婦服地黄湯來，似減。 眠食尚未佳，憂懸 不去心。君等前所論事， 想必及。謝生未還，可爾。 進退不可解，吾當書問也。
	靈堂永晝，來風長 石枕。竹簟生清光， 熱文園肺渴。 厭煩熱，更要夫君 在側，傍	虛堂永晝來風長， 石枕竹簟生清光。 文園肺渴厭煩熱， 更要夫君在側傍。	虛堂永晝來風長， 石枕竹簟生清光。 文園肺渴厭煩熱， 更要夫君在側傍。

Table 5: Cases studies.

Secondly, our focus has been primarily on a single language. While we have achieved promising results in this language, it is important to acknowledge that the generalizability of our approach is limited since other languages may not have the enough calligraphy work.

Ethical Statement

For the annotating our CCR dataset, we hired 10 annotators and tasked 20 works each person, with a payment of 19 CNY for each calligraphy work. The work was down within 3 hours so their average hourly wage was higher than 100 CNY; For the human recognition, we hired 10 annotators and tasked 20 works each person, with a payment of 3 CNY for each calligraphy work. The work was down within 2 hours so their average hourly wage was higher than 30 CNY; Both payments were far higher than the local low hourly wage standard (19 CNY per hour).

Acknowledgments

We would like to thank the anonymous reviewers for their excellent feedback. This work is supported by The Hong Kong Polytechnic University Projects (#P0048932, #P0051089, #P0056456), the National Natural Science Foundation of China (No. 62376178), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Antonis Anastasopoulos and David Chiang. 2018. Leveraging translations for speech transcription in low-resource settings.

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.

Xiaoyi Bao, Jinghang Gu, Zhongqing Wang, Minjie Qiang, and Chu-Ren Huang. 2024. Employing glyphic information for Chinese event extraction with vision-language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1068–1080, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoyi Bao, Zhongqing Wang, Jinghang Gu, and Chu-Ren Huang. 2025. Revisiting classical Chinese event extraction with ancient literature information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8440–8451, Vienna, Austria. 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.

Jingye Chen, Bin Li, and Xiangyang Xue. 2021. Zero-shot chinese character recognition with stroke-level decomposition.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites.

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.

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, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model.

Senka Drobac, Pekka Kauppinen, and Krister Lindén. 2017. OCR and post-correction of historical Finnish texts. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 70–76, Gothenburg, Sweden. Association for Computational Linguistics.

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.

Ji-dan Huang, Guanjie Cheng, Jinghan Zhang, and Wei Miao. 2022. Recognition method for stone carved calligraphy characters based on a convolutional neural network. *Neural Computing and Applications*, 35:1–10.

JaidedAI. Easyocr.

Amrith Krishna, Bodhisattwa P. Majumder, Rajesh Bhat, and Pawan Goyal. 2018. Upcycle your OCR: Reusing OCRs for post-OCR text correction in Romanised Sanskrit. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 345–355, Brussels, Belgium. Association for Computational Linguistics.

Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. 2016. Textboxes: A fast text detector with a single deep neural network.

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

Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiufeng Wang. 2013. Online and offline handwritten chinese character recognition: Benchmarking on new databases. *Pattern Recognition*, 46:155–162.

Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024. Focus anywhere for fine-grained multi-page document understanding.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

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.

OpenAI. 2024. Gpt-4o system card.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

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.

Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored OCR post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2726, Copenhagen, Denmark. 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.

Kaili Wang, Yaohua Yi, Junjie Liu, Liqiong Lu, and Ying Song. 2020. Multi-scene ancient chinese text recognition. *Neurocomputing*, 377:64–72.

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*.

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.

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.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. 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, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding.

Yue Xu, Fei Yin, Da-Han Wang, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. 2019. Casia-ahcdb: A large-scale chinese ancient handwritten characters database. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 793–798.

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, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 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 Performance on Standard Font Image

As our model is tested on historic calligraphy works, we further check if our model is effective in standard font images, thereby provides a glimpse of our model's ability to recognize common OCR scenarios, where the characters are usually in printed standard font. Specifically, we collect extra 300 images under the same criteria of building our CCR trainset, and use it as the testset to test model's performance in standard fonts.

From Table 6, we can tell that even on standard font images, our model still outperform the baselines with a slight margin. This underscores our model not only specialize in calligraphy recognition, but also generalize to common OCR situations. Moreover, all the baselines perform relatively much better on standard font images than historic calligraphy, indicating that our end-to-end calligraphy recognition is a difficult task compared to the common OCR task.

Method	**↑ F1.**	**↓ CER**	**↑ BLEU**
Traditional OCR Baselines			
PaddleOCR(off-the-shelf)	0.9172	0.0856	0.7962
EasyOCR(off-the-shelf)	0.8935	0.0921	0.7772
EffOCR	0.9363	0.0873	0.8071
VLM-driven OCR Baselines			
GOT-OCR2.0	0.8745	0.1067	0.7314
Vary	0.8234	0.1678	0.6023
LLaVA-1.5-7B	0.7738	0.1743	0.5065
Qwen2-VL-7B	0.9423	0.0891	0.7529
Qwen2.5-VL-7B	0.9493	0.0699	0.7854
InternLM-XComposer	0.9147	0.0767	0.7731
InternVL2.5-8B	0.9544	0.0734	0.7945
Ours	**0.9772**	**0.0509**	**0.8411**

Table 6: Performance on standard font.

Formation		**↑ F1.**	**↓ CER**
Input Image	**Output Target**		
Simplified	Simplified	0.6174	0.6179
Simplified	Traditional	0.6021	0.6343
Traditional	Simplified	0.5934	0.6431
Traditional	Traditional	**0.6715**	**0.5318**

Table 7: The impact of Chinese formations.

B Impact of Formations

The formation of the input image and output target during training, whether traditional or simplified, could be vital to the final recognition. Although the traditional formation ensures the consistency throughout the entire training and inference process, it is not well-suited for language models, which are primarily pretrained on corpora where simplified Chinese is the dominant language. We thus investigate the impact of formation on our recognition task.

From Table 7 we can tell that, both the two consistent pairs outperform the inconsistent, which is excepted since the inconsistent formation will cause a fissure in semantic understanding. Moreover, among two consistent pairs, the traditional surpass the simplified, which gives us the conclusion that the consistency of formation throughout the modeling is more crucial to the recognition and the deficiency in semantic understanding can be remedied by downstream finetuning.

C More Cases

We add more case studies in Table 8. These cases demonstrate the versatility of our model in adapting to different input styles. Specifically, the first two examples highlight the model's robustness in interpreting and processing freehand scribbles, while the last three examples showcase its ability to produce high-quality outputs from cleaner and more structured inputs.

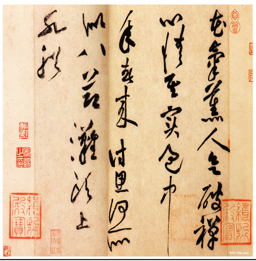
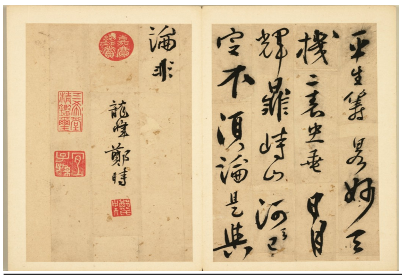
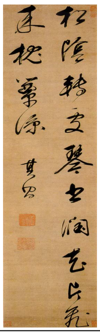
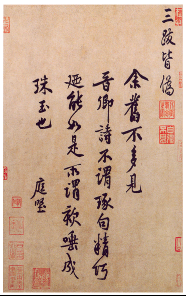
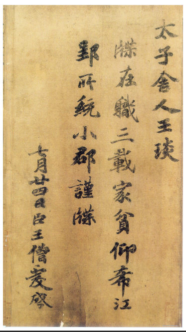
Input Calligraphy Image	Ours	Ground Truth
	花氣薰人欲破禅， 唯將至實包中年。 年克森來忖思何似， 公節灘頭上水船。	花氣薰人欲破禅， 心情其實過中年。 春來詩思何所似， 八節灘頭上水船。
	平生籌略妙天機， 二表忠垂日月輝， 鼎鼎峙山天已定， 河漢空不須論是。	平生籌略妙天機， 二表忠垂日月輝。 鼎峙山河天已定， 不須論是與論非。
	松陰轉處琴書潤， 花片飛來枕簟涼。	松陰轉處琴書潤， 花片飛來枕簟涼。
	余舊不多見晋卿詩， 不謂琢句精 仍能如是。 所謂亥欠唾成珠玉也。	余舊多不見晋卿詩， 不謂琢句精巧， 乃能如是， 所謂亥欠唾成珠玉也。
	太子舍人王琰牒， 在職三載，家貧 仰希江覢所統小郡， 謹牒。	太子舍人王琰牒。 在職三載，家貧， 仰希江郢所統小郡， 謹牒。

Table 8: More Cases.