EvaHan 2026 Ancient Books Multimodal OCR and Layout Analysis System Technical Report

Chenrui Zheng
Department of Chinese, Sun Yat-sen University
zhengchr5@mail2.sysu.edu.cn

Abstract

This paper introduces our system proposal and experimental results for the 5th International Evaluation of Ancient Chinese Information Processing (EvaHan 2026). This evaluation focuses on ancient books OCR tasks using multimodal large language models, including three subtasks: Printed Text Recognition (Task A), Layout Element Analysis (Task B), and Handwritten Text Recognition (Task C). To address core challenges such as numerous variant characters, complex handwritten ligatures, dense layout elements, and annotation noise, we propose a Supervised Fine-tuning (SFT) scheme based on data synthesis augmentation and multi-stage curriculum learning. We also optimized the data preprocessing workflow, resolving key issues like repetition mark recognition and annotation quality improvement. We completed a 9:1 train-validation split on the official dataset and verified the effectiveness of our methods through 6 groups of comparative experiments. Finally, we selected the model with the best comprehensive performance for submission. The code and synthetic dataset are available at https://github.com/zhengningch/EvaHan2026-data.

Keywords: Ancient Books OCR,Curriculum Learning, Data Augmentation

1. Introduction

1.1. Background and Significance

Ancient books are the core cultural heritage of Chinese civilization. Existing ancient books are mostly preserved in the form of manuscripts or ancient printed editions, which are fragile and difficult to disseminate on a large scale. Optical Character Recognition (OCR) technology can transform paper/image-based ancient books into editable digital text, serving as a core foundational technology for digital preservation, intelligent retrieval, and humanities research. Compared to modern text OCR, ancient books OCR faces multiple challenges such as variant characters, handwritten ligatures, complex layouts, and stain occlusions. The development of multimodal large language models (MLLMs) provides new solutions for end-to-end ancient books OCR and layout analysis.

1.2. Task Introduction

EvaHan 2026 is the 5th International Evaluation of Ancient Chinese Information Processing, focusing on multimodal ancient books OCR tasks. It consists of three core subtasks: **Task A: Printed Text Recognition**, aiming to output Traditional Chinese text from printed page images, evaluated by Character Error Rate (CER); **Task B: Layout Element Analysis**, requiring the detection of text, image, book_edge, and seal elements with bounding boxes, evaluated by mAP@[.5:.95]; and **Task C: Handwritten Character Recognition**, targeting handwritten text recognition, also evaluated by CER. The challenges include dense variant characters, complex layouts, stain occlusions, and personalized handwritten styles. All experiments strictly follow the Closed Modality constraint using the Qwen2.5-VL-7B-Instruct model. As a baseline, we applied conventional SFT using the official training set without data augmentation or multi-stage training.

2. Related Work

2.1. OCR Technology Research

OCR technology has evolved from traditional machine learning methods (such as Support Vector Machines) to deep learning methods. Document understanding tasks have integrated these deep learning models into a modular pipeline. To achieve document understanding of complex layouts, it is necessary to sequentially perform stages such as layout analysis (identifying spatial coordinates and reading order of document elements), content extraction (text OCR, mathematical formulas converted to LaTeX format, etc.), and relationship integration. For example, Baidu's PP-StructureV2 and PP-StructureV3 have integrated modular architectures for multi-task collaboration including data preprocessing, OCR modules, layout analysis, and document element recognition (Baidu, 2024); Wang et al. (2024) similarly integrates the PDF-Extract-Kit model library (containing layout detection, formula detection/recognition, table recognition, OCR, etc.) to achieve high-quality extraction of different documents. This applies to document understanding in the ancient books domain as well. For instance, Saini et al. (2019) achieved first place in three tasks at the

Proceedings of LT4HALA 2026 @ LREC 2026, pages 288–293.
11 May 2026.

ICDAR 2019 Large Structured Chinese Family Records Reading Challenge by selecting appropriate architectures for detection, segmentation, and recognition modules and achieving collaborative optimization.

2.2. Layout Analysis Research

The rise and maturation of Vision Language Models (VLMs) have brought a paradigm shift to document understanding. Many R&D teams have shifted their focus to end-to-end VLMs. For example, Niu et al. (2025) released MinerU2.5, improved based on the Qwen2-VL framework to achieve SOTA performance while maintaining significantly lower computational costs. Another example is olmOCR launched by the Allen Institute for AI (Poznanski et al., 2025), which is based on a 7B VLM fine-tuned from Qwen2-VL-7B-Instruct, improving output quality through Document-Anchoring technology (combining PDF text blocks with position information). On the other hand, Cui et al. (2025) released PaddleOCR-VL, which adopts a hybrid system using a lightweight VLM while first adding PP-DocLayoutV2 to ensure the text order conforms to human reading logic. Utilizing VLMs to understand complex documents has great potential. However, current research mainly focuses on modern books (such as academic papers, math exams, mixed image-text manuals). For the specific domain of ancient books, Li et al. (2025) pointed out that "existing large models still struggle to accurately read text content in order and locate it when processing relatively clear ancient book documents". Therefore, it is necessary to attempt to explore the potential of VLMs in the vertical domain of ancient books.

2.3. Application of Curriculum Learning

Ancient book layout analysis involves detecting diverse elements with varying degrees of complexity, from simple text blocks to overlapping seals. To effectively handle these challenges, we adopt a curriculum learning approach. Curriculum Learning (CL) was originally proposed by Bengio et al. (2009), imitating the human learning process where training starts with simple examples and gradually transitions to more complex ones. This strategy helps models find better local optima in non-convex optimization spaces and accelerates convergence, as highlighted in recent surveys (Wang et al., 2021; Soviany et al., 2022). In the field of Optical Character Recognition (OCR), CL has demonstrated significant benefits. For instance, a work by Borisyuk et al. (2018) utilized CL in Facebook's Rosetta system by progressively increasing the difficulty of training data (e.g., text length, deformation degree), which significantly improved the performance and robustness of their sequence recognition models. Similarly, in Handwritten Text Recognition (HTR), CL has been effectively applied to handle variable sequence lengths and noise in historical documents (Wang et al., 2021), improving both convergence speed and final accuracy. Given the complexity of ancient Chinese characters and layouts, we adopt a multi-stage CL strategy to stabilize the fine-tuning of our multimodal model.

3. Methodology

This section details our system proposal, including three core modules: data preprocessing, data synthesis and augmentation, and multi-stage curriculum learning fine-tuning strategy.

3.1. Data Preprocessing and Dataset Construction

3.1.1. Dataset Split

The three datasets provided officially each contain about 5000 image-text pairs. We divided each dataset into training and validation sets at a 9:1 ratio, i.e., 4500 samples for training and 500 samples for local validation per subset, ensuring distribution consistency between training and validation data.

3.1.2. Preprocessing for Task A and Task C

To improve model focus on character recognition and address the challenge of repetition marks in Task C, we designed a specialized preprocessing and post-processing scheme. First, we removed all punctuation marks from the ground truth text for both Task A and Task C to reduce noise and focus the model on character recognition. Second, we replaced the positions of repetition marks in the training set with a special token $\star$ to reduce learning difficulty. Finally, using regularization rules, we automatically restored the $\star$ output by the model to the corresponding preceding Chinese character to ensure accuracy.

3.1.3. Annotation Optimization for Task B

During data verification, we found some annotation errors in the official Task B training set, mainly including missed elements, incorrect labels, and BBOX coordinate deviations. Therefore, we performed manual secondary verification and correction on all Task B training data: supplementing missed layout elements, correcting wrong labels and coordinates (converting four-point coordinates to bbox format), and removing 480 images with

overly complex layouts or unverifiable annotation quality, significantly improving training data quality. See Figure 1 for examples of our data preprocessing and correction efforts.

3.2. Data Synthesis and Augmentation Strategy

To address the issues of insufficient diversity and low proportion of hard samples in ancient book training data, we designed adapted synthesis and augmentation schemes for the three subtasks. See Figure 2 for an overview of the workflow.

3.2.1. Data Augmentation for Task A and Task C

1. **Hard Sample Synthesis:** We screened 1000 complex images with high recognition difficulty from A and C training sets. Based on pixel points, we cut them into single-character images, called the Qwen2.5-VL-72B model for single-character OCR recognition, and filtered out single-character samples with recognition errors. We randomly combined the error single-character images with correct labels, generating 1000 new synthetic training data entries following A and C styles to reinforce the model's ability to recognize difficult characters.

2. **General Data Augmentation:** We performed random blurring, stroke thickening/thinning, stain addition, and random cropping on the original training set to expand data diversity and improve model robustness against low-quality images.

3.2.2. Data Synthesis and Augmentation for Task B

Based on annotated BBOX coordinates, we first built a layout element material library by fine-grained cutting of text and seal elements. Then, we performed layout reorganization and synthesis with the following strategies: (1) **Edge Constraint:** ensuring book_edge elements are strictly located at the sides to simulate binding features; (2) **Random Distribution:** non-overlapping random filling of elements to simulate diverse layouts such as Horizontal/Vertical Multi-grid, Circular/Ring Text Arrangement, Image-Text Mixing, and Random Arrangement; (3) **Scale Expansion:** generating 2500 high-quality synthetic images with pixel sizes consistent with the original set. Finally, we applied general data augmentation (Flip, Blur, Contrast) to basic training data to prevent overfitting.

3.3. Multi-stage Curriculum Learning Fine-tuning Strategy

We used the officially designated Qwen2.5-VL-7B-Instruct as the base model, adopting LoRA fine-tuning as the primary method and full-parameter fine-tuning as auxiliary. Based on the core concept of curriculum learning "from simple to difficult", we designed a multi-stage fine-tuning scheme. To avoid catastrophic forgetting, 500-700 basic data entries from A and C tasks were added in each training stage.

3.3.1. Fine-tuning Process for Task A and Task C

Stage 1 Basic SFT: Use the original training set and basic augmented data for conventional LoRA fine-tuning to obtain a baseline model, allowing the model to master basic ancient book OCR task formats and recognition capabilities. **Stage 2 Hard Sample Optimization:** For complex samples, handwritten ligatures, and repetition mark samples with poor recognition in Stage 1, add synthetic hard samples and augmented data for a second round of LoRA fine-tuning.

3.3.2. Difficulty Clustering for Task B Curriculum Learning

To implement the simple-to-complex Curriculum Learning strategy, we performed K-Means clustering analysis on the cleaned Task B training set based on element Count, Density, and Overlap, dividing all samples into three difficulty levels. **Easy (2400 entries):** Sparse layout elements (< 5 bbox), clear layout, no overlap. Used for fast convergence in early training. **Medium (949 entries):** Moderate number of elements, minor adjacency or slight occlusion. Used to consolidate boundary regression ability. **Hard (723 entries):** Extremely dense elements (> 50 bbox), significant seal/text overlap, tight text line arrangement. Used to tackle missed/false detections in complex layouts.

3.3.3. Multi-group Fine-tuning Experiments for Task B

Addressing the high difficulty and poor baseline of Task B, we designed 6 groups of comparative experiments:

1. **Stage1-LoRA:** Only use simple and synthetic data for LoRA fine-tuning.

2. **Stage1-LoRA + Stage2-LoRA:** Add medium difficulty and augmented data for a second round.

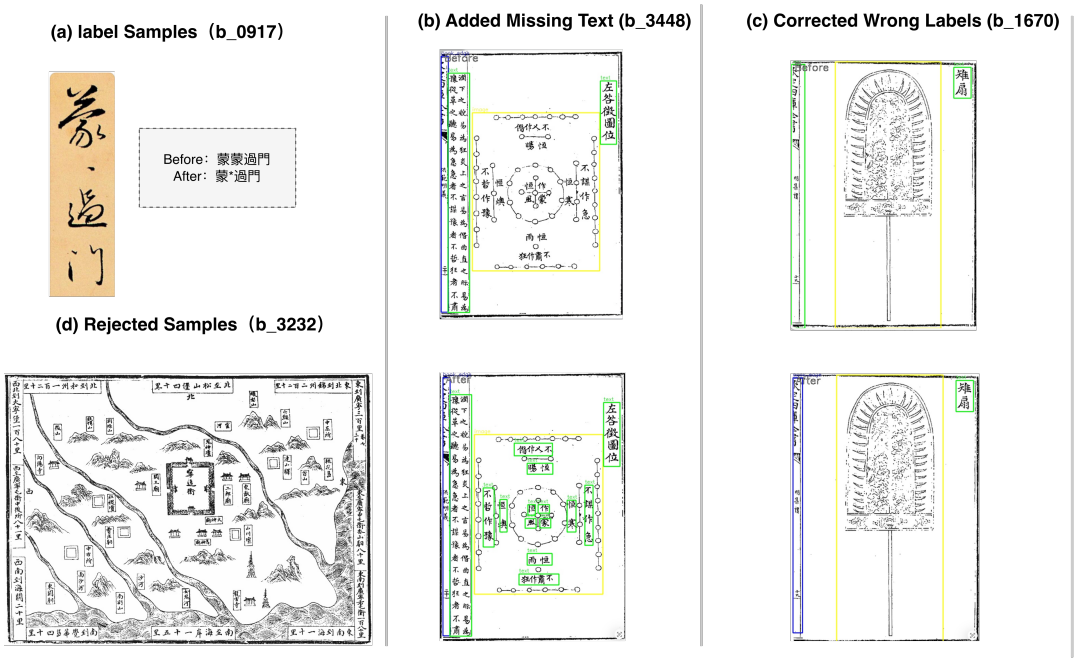

Figure 1: Data Preprocessing Examples

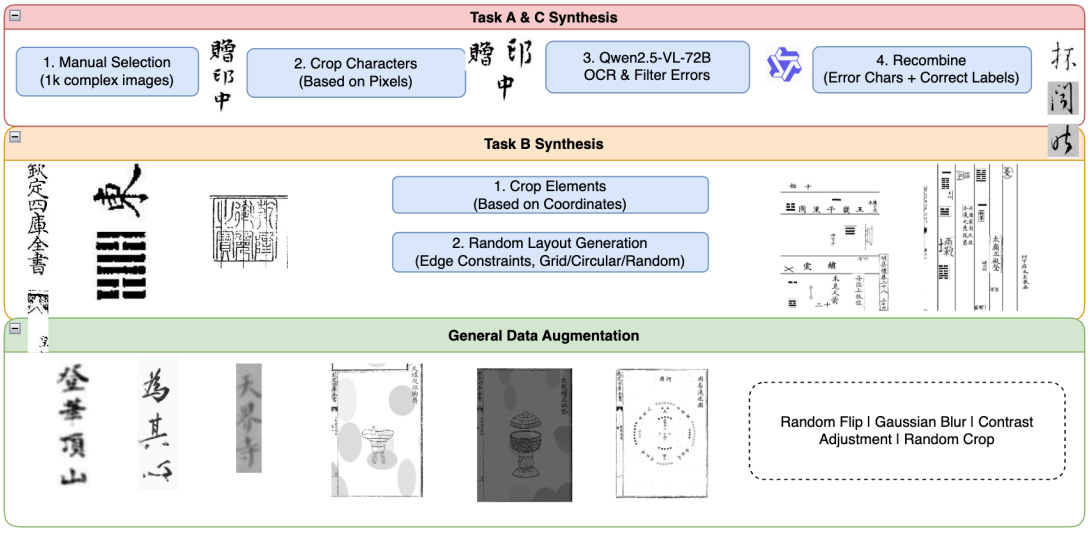

Figure 2: Overview of our Data Synthesis and Augmentation Pipeline.

3. **Stage1-LoRA + Stage2-LoRA + Stage3-LoRA:** Add remaining medium and hard data for a third round.

4. **Stage1-LoRA + Stage2-Full + Stage3-Full:** Perform full-parameter fine-tuning for medium and hard data respectively.

5. **Stage1-LoRA + Stage2-LoRA + Stage3-GRPO:** Use GRPO reinforcement learning for hard data. We designed a composite reward function: (1) **OCR Accuracy Reward:** Combination of mAP and Soft-IoU for box precision, with a difficulty bonus for dense layouts ($R_{ocr} = (0.5 \cdot \text{mAP} + 0.5 \cdot \text{DenseScore}) \times (1 + 0.1 \ln(\text{count}))$); (2) **Format Reward:** +0.2 for valid JSON and XML tags; (3) **Repetition Penalty:** -1.0 for detected N-gram loops or duplicate BBoxes.

6. **Stage1-LoRA + Stage2-GRPO:** Use GRPO for medium data.

4. Experiments

4.1. Experimental Setup

We implemented all experiments based on the ms-swift framework. The core hyperparameters are: learning rate = 1e-5, batch size = 8, max sequence

length = 4096, epochs = 1. The LoRA configuration is: rank=64, alpha=128, dropout=0.05, target modules: visual encoder and LLM attention layers. For GRPO experiments, we set learning rate to 1e-6, number of generations to 8, and used vLLM for accelerated rollout with a repetition penalty of 1.05. The hardware environment is: 4090×2 / H10 $\times 1$ GPU.

4.2. Evaluation Metrics and Baseline Model

We strictly followed the official evaluation standards: Task A, C: Main metric CER (lower is better); auxiliary metrics Character-level F1, NED.Task B: Main metric mAP@[.5:.95] (higher is better); auxiliary metrics IoU, Micro-average F1. We used qwen2.5-vl-7b-instruct and the official training set for conventional SFT without extra data augmentation or multi-stage training as the baseline model.

4.3. Validation Set Results

Table 1 and Figure 3 summarize our validation results.

Model	Task A	Task C
Baseline	10.0	22.0
Stage 1	**2.0**	14.0
Stage 2	3.0	**10.0**

Table 1: Task A and Task C Validation Set Performance (CER %)

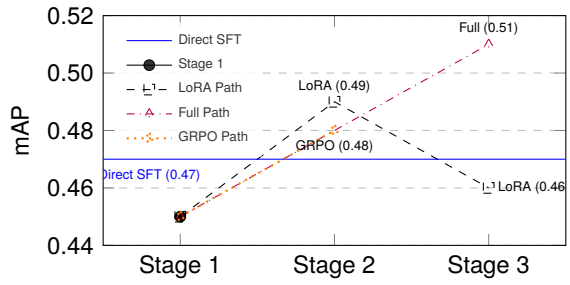

Figure 3: Task B mAP Performance Progression. Direct SFT (0.47) is shown as a reference line. Baseline performance (0.09) is omitted from the plot for clarity due to scale differences. The plot highlights the improvement in Stage 2 and 3 compared to Stage 1 (0.45).

5. Results and Discussion

5.1. Core Experimental Conclusions

Curriculum Learning SFT outperforms conventional SFT: The simple-to-difficult approach prevents overfitting on simple samples and improves capability on hard samples. Task B mAP improved by 10.9 percentage points over baseline. **Full-parameter fine-tuning shows slight improvement:** Compared to LoRA, full-parameter fine-tuning improved mAP slightly but not significantly. **Hard data can degrade performance:** Forcing extremely hard samples in the SFT phase reduced robustness and triggered N-gram repetition errors. **GRPO did not meet expectations:** GRPO failed to improve performance and instead led to severe instability. We hypothesize that training on extremely complex samples actually deteriorates model robustness. This is likely because the dense, repetitive layout elements (BBox coordinates) force the model into N-gram repetition loops. Under such an unstable environment with small inter-group variance, the 7B base model struggles to learn meaningful new policies. This remains an open challenge for future work.

5.2. Final Submitted Model

We selected **Stage1-LoRA + Stage2-Full + Stage3-Full** as the submission model, achieving optimal comprehensive performance and stability on the validation set.

6. Conclusion

This paper proposed a fine-tuning scheme based on data synthesis augmentation and multi-stage curriculum learning for EvaHan 2026. We optimized preprocessing and solved issues like repetition marks and annotation noise. Comparative experiments verified the effectiveness of phased curriculum learning. We also explored full-parameter fine-tuning and GRPO. Additionally, performance on handwritten text (Task C) indicates room for improvement, specifically regarding cursive stroke adhesion and complex variant characters. Future work will address these limitations by designing joint optimization frameworks and exploring reinforcement learning strategies more suitable for multimodal ancient book tasks.

7. Acknowledgements

We thank the EvaHan 2026 organizing committee for the high-quality datasets and platform, and LREC 2026 / LT4HALA 2026 workshop for academic exchange opportunities.

8. Bibliographical References

Baidu. PP-Structure: Document Structure Analysis. https://github.com/PaddlePaddle/PaddleOCR, 2024.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pp. 41-48, 2009.

F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 71-79, 2018.

C. Cui, T. Sun, M. Lin, et al., "PaddleOCR 3.0 Technical Report," arXiv preprint arXiv:2507.05595v1, 2025.

H. Li, Y. Liu, W. Liao, et al., "Optical Character Recognition in the Era of Large Models: Status and Prospect," *Journal of Image and Graphics*, vol. 30, no. 6, pp. 2023-2050, 2025.

B. Wang, C. Xu, X. Zhao, et al., "MinerU: An Open-Source Solution for Precise Document Content Extraction," arXiv:2409.18839, 2024.

J. Niu, Z. Liu, Z. Gu, et al., "MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing," arXiv:2509.22186, 2025.

J. Poznanski, A. Rangapur, J. Borchardt, et al., "olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models," arXiv:2502.18443, 2025.

R. Saini et al., "ICDAR 2019 Historical Document Reading Challenge on Large Structured Chinese Family Records," in *IEEE International Conference on Document Analysis and Recognition (ICDAR)*, 2019.

P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum Learning: A Survey," *International Journal of Computer Vision*, vol. 130, pp. 877–902, 2022.

X. Wang, Y. Chen, and W. Zhu, "A Survey on Curriculum Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4555-4576, 2021.