Beijing Normal University at EvaHan 2026: Enhancing Ancient Chinese Character Recognition and Layout Analysis via VLM Fine-Tuning and Linguistic Post-Processing

Yihuan Yin[1], **Qian Zhao**[1]
[1]Institute of Chinese Information Processing,
School of International Chinese Language Education, Beijing Normal University
{202421091021, 202421091022}@mail.bnu.edu.cn

Abstract

This paper describes the system submitted by the Beijing Normal University (BNU) team for the EvaHan 2026 shared task. We participated in Task A (Printed Text Recognition), Task B (Layout Element Analysis), and Task C (Handwritten Character Recognition). For text recognition (Tasks A and C), we proposed a hybrid pipeline combining supervised fine-tuning (SFT) of Vision-Language Models (VLMs) with a linguistic rule-based post-processing module. In the Open Track, we further explored the use of a general-purpose VLM to correct semantic errors while maintaining visual fidelity to ancient variant characters. For Task B, we adopted a method integrating a VLM with structured prompting strategies. Our system consistently surpassed the official baselines, achieving an F1 score of 94.53% in Task A and 91.33% in Task C, while demonstrating enhanced localization precision in Task B.

Keywords: Ancient Chinese, OCR, Layout Analysis, Vision-Language Models, LoRA

1. Introduction

The digitization of ancient Chinese literature poses significant challenges due to the presence of variant characters, complex layouts, and idiosyncratic handwritten cursive styles. The EvaHan 2026 shared task provides a comprehensive platform to evaluate multimodal systems on historical documents across three tracks: Task A, Task B, and Task C.

Our team from BNU approached this challenge by leveraging the recent advancements in multimodal large language models (MLLMs). For text recognition, we focused on addressing the ”hallucination” and ”over-standardization” issues where models arbitrarily convert ancient variant characters into modern standard forms. For layout analysis (Task B), we concentrated on multi-category structural detection in mixed image-text pages under the closed-track setting. Without using any external data, our system incorporates domain-specific fine-tuning and linguistic deterministic rules to ensure both semantic accuracy and structural validity.

2. Related Works

Document layout analysis has evolved from traditional image processing techniques to deep neural network–based approaches. Early methods relied on projection profiles and connected component analysis for page segmentation (Nagy, 2000), achieving reasonable performance in documents with regular layouts. With the development of deep learning, general object detection architectures such as Faster R-CNN and Mask R-CNN demonstrated stronger performance in document region detection tasks (Ren et al., 2015; He et al., 2017).

To further integrate textual content and spatial layout information, Transformer-based multimodal models such as LayoutLM (Xu et al., 2020) jointly model OCR text, positional embeddings, and visual features, significantly improving performance. However, these methods are primarily designed for modern printed documents. The diversity and noise characteristics of ancient documents limit the direct transferability of such models. Studies on historical document analysis have introduced U-Net–based segmentation frameworks (Oliveira et al., 2018) and multi-scale feature fusion strategies, but multi-category layout detection with structured output remains challenging.

Furthermore, while VLMs demonstrate strong multimodal alignment capabilities (Bai et al., 2023), their application combined with parameter-efficient fine-tuning (e.g., LoRA (Hu et al., 2022)) and structured post-processing in ancient book layout detection remains relatively underexplored. This study aims to bridge this gap, aligning with the evaluation objectives of the EvaHan community (Li et al., 2024).

3. Methodology

Our system processes inputs through a structured pipeline depending on the task and evaluation track.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 273–276.

3.1. Text Recognition (Tasks A and C)

In the Closed Track, we utilized the Xunzi-Qwen2-VL-7B-Instruct model (Wang et al., 2023). To mitigate structural OCR errors, we implemented a linguistic deterministic rule-based post-processing module. This included merging split components and fixing context-specific terminology errors based on historical official titles.

For the Open Track, we employed Doubao-1.5-Pro (Vision) as a semantic correction agent with a "Conservative Correction Strategy." The prompt explicitly instructed the model to preserve handwritten variants to maintain fidelity to the original manuscripts, while correcting genuine semantic errors based on historical contexts.

3.2. Layout Element Analysis (Task B)

Task B focuses on layout element recognition. The core objective is the precise identification of four key elements within the pages of ancient books: text, image, book_edge, and seal. Under the closed-track setting, all participating systems are required to use the officially specified models. In our implementation, we adopted Qwen2.5-VL-7B-Instruct as the backbone model. Based on this architecture, we constructed a layout detection framework and enhanced its adaptability to ancient book layout analysis through parameter-efficient fine-tuning using LoRA.

3.3. Implementation Details

To ensure reproducibility and comprehensively address the evaluation criteria, we detail the hyperparameter configurations, hardware environment, and post-processing strategies below.

Parameter / Component	Configuration / Version
Fine-tuning Method	LoRA (Scope: all linear layers)
LoRA Alpha / Dropout	16 / 0.0
Learning Rate	2×10^{-4} (Cosine Scheduler)
Warmup / Weight Decay	0 steps / 0.0
Max Grad Norm	1.0
Epochs / Precision	3 / `fp16` mixed precision
Effective Batch Size	8 (Per-device: 1, Accumulation: 8)
Hardware Configuration	NVIDIA GeForce RTX 4090 D $\times$ 1
Peak VRAM Consumption	$\sim$20–22 GB
CPU / Memory	15 Cores / 80 GB
Software Environment	Python 3.11.14, CUDA 13.0
Core Libraries	PyTorch 2.6.0+cu124
	Transformers 4.57.1, vLLM 0.4.0
	LLaMA-Factory 0.9.5.dev0

Table 1: System Environment and Fine-Tuning Hyperparameters for VLM Adaptation.

VLM Fine-Tuning Configurations: As detailed in Table 1, the core visual-language models were fine-tuned using the LLaMA-Factory framework. Although `bf16` is often preferred for fine-tuning Qwen2.5-VL due to its wider numerical range and better stability, we utilized `fp16` mixed precision based on our specific hardware constraints and environment compatibility. To mitigate potential gradient overflow issues inherent to `fp16`, we applied strict gradient clipping (max norm $= 1.0$) and dynamic loss scaling. In practice, no gradient overflow or obvious numerical instability was observed, and the model completed training stably across all epochs.

Open Track Prompting Strategy: For generative post-correction in the open track, we designed track-specific prompts to constrain the VLM. In Task A, the prompt enforced a "visual fidelity" rule, strictly prohibiting the conversion of historical colloquial characters (e.g., 万, 无) into standard traditional forms unless visually justified. In Task C, the prompt instructed the model to trust the base OCR for rare entities while forcing it to correct financial logic and numeral errors based on visual evidence.

Closed Track Linguistic Post-Processing: We introduced a deterministic linguistic post-processing module involving a dual-step pipeline: (1) Aggressive format cleaning using regular expressions to strip out all punctuation and full-width spaces. (2) Hard-coded symbolic overrides tailored to specific historical contexts, correcting high-confidence misrecognitions related to official titles and financial terms (e.g., 大住 to 大臣) via a predefined expert lexicon. Crucially, these rules were not arbitrary; they were systematically derived through a manual error analysis of the base model's initial predictions, deeply integrated with domain-specific knowledge of ancient Chinese culture and administrative history.

4. Experiments and Results

4.1. Dataset

The dataset provided by the EvaHan 2026 organizers contains mixed image-text data selected from the Siku Quanshu and other ancient books. Each sample consists of an ancient book page image and the corresponding official annotation in JSON format. The annotations define four structural entity categories (text, image, seal, book_edge), with each entity including its category label and four coordinate points ordered as top-left, top-right, bottom-right, and bottom-left.

4.2. Main Results

The quantitative evaluation of our proposed system on the EvaHan 2026 test set is presented in Table 2.

Track & System	Task A F1 (%)↑	Task B Macro F1 (%)↑	Task C F1 (%)↑
Official Baseline	94.30	**15.30**	90.99
Our System (Closed)	**94.53**	11.73	**91.33**
Our System (Open)	-	-	-

Table 2: Performance comparison on the EvaHan 2026 test set. F1 (%) indicates the F1 score in percentage. (Note: Official quantitative results for the Open Track are currently pending; thus, our Open Track submission is discussed primarily as a qualitative exploration of VLM capabilities.)

Text Recognition (Tasks A and C): The official baseline for Task A (Printed Texts) yielded an F1 score of 94.30%. Our system achieved an F1 score of 94.53%, outperforming the baseline. For Task C (Handwritten Texts), our proposed pipeline achieved an F1 score of 91.33%, surpassing the baseline's 90.99%, demonstrating robustness on highly personalized cursive connections.

Layout Element Analysis (Task B): Under the closed-track setting, our system achieved an Avg Match IoU of 0.7048, a Micro F1 of 7.14%, and a Macro F1 of 11.73%. As shown in Table 2, while the official baseline achieved a higher Macro F1 of 15.30% and a higher mAP@[.5:.95] (0.2006 vs. our 0.1402), our system significantly improved the average matching IoU (from 0.6600 to 0.7048). This indicates more precise bounding box localization for detected elements, albeit with a reduced overall detection performance.

4.3. Ablation Study and Error Analysis

To verify the contributions of our proposed modules and analyze the performance bottlenecks mentioned above, we conducted qualitative ablation studies and error analyses across the tasks.

Impact of Linguistic Post-Processing and Prompting (Tasks A & C): Removing the linguistic post-processing module resulted in a noticeable resurgence of deterministic terminology errors (e.g., historical official titles being misclassified due to modern linguistic priors). The structured prompting strategy was equally critical; without the "visual fidelity" constraints, the VLM exhibited severe normalization hallucination, arbitrarily converting historical variant characters into standardized modern forms and disrupting the original layout formatting. These two modules were the primary drivers behind our consistent F1 improvements over the baseline.

Small Object Detection Bottleneck (Task B): Our layout analysis pipeline currently lacks targeted designs for small object detection, which directly explains the severe long-tailed class imbalance. Specifically, the seal category suffered from near-zero recall. Seals in ancient texts are inherently minority classes, often appearing as small, faded, and low-contrast regions. Because our generic VLM fine-tuning approach did not incorporate specialized small-object mechanisms (such as multi-scale feature fusion or dynamic anchor oversampling), the model prioritized high-frequency, large-area objects (like text blocks) to minimize the global loss. This systematic degradation in minority categories significantly pulled down the overall Macro F1 score, despite the high localization precision (IoU) achieved on text elements.

5. Discussion

The experimental results for Task B demonstrate a clear improvement in spatial localization capability but reveal a structural imbalance at the category level. The increase in IoU indicates that LoRA fine-tuning effectively enhances the model's geometric modeling of page structure, producing tighter and more accurate bounding boxes.

As analyzed in the ablation study, further manual inspection reveals that VLMs tend to adopt a conservative prediction strategy, refraining from outputting detections when confidence is insufficient in order to reduce false positives. While this strategy improves localization precision for large objects, it results in systematic recall loss for small objects like seals. Moreover, minimum-size filtering rules may further suppress seal detections, as small bounding boxes shrink excessively during discretization. Therefore, the system exhibits a characteristic pattern of "high localization precision but insufficient category coverage." If seal recall can be improved while maintaining the current IoU level, both Macro F1 and mAP are expected to increase.

6. Conclusion

We successfully developed a hybrid VLM-based pipeline for ancient Chinese texts. By combining LoRA fine-tuning with targeted linguistic post-processing, our system surpassed the official baselines in character recognition tasks. For layout analysis, utilizing structured prompts and post-processing constraints yielded tighter bounding box localization. Future work will explore integrating Retrieval-Augmented Generation (RAG) and specialized small-object detection strategies to improve minority category recall.

7. Bibliographical References

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

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Bin Li et al. 2024. Overview of evahan 2024: The evaluation of ancient chinese language processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

George Nagy. 2000. Twenty years of document image analysis in pami. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38–62.

Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. 2018. dhsegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 7–12. IEEE.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Dongbo Wang et al. 2023. Xunzi: Open-source large language model for chinese ancient text processing. Model open-sourced on GitHub and ModelScope.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.