LVLM Optimization for Ancient Chinese Book Image Analysis with Task-specific Augmentation and Instruction Tuning

Xia Tian[1], Liu Yulong[2], Wang Yilin[1], Yang Yumeng[1]
Cai Dongheng[1], Tan Yuyang[3],Yang Menghui[1]*
[1]School of Information Resource Management,Renmin University of China, Beijing, China,
[2]Midu Technology Co., Ltd., Shanghai,[3]China Washington University in St. Louis
xiat@ruc.edu.cn, lcayvinl@gmail.com, wang_yilin@ruc.edu.cn,
{yymeng1219,dongheng_c}@163.com,tan@wustl.edu,yangmenghui@ruc.edu.cn

Abstract

Ancient Chinese text digitization faces challenges like variant characters and complex layouts. Based on the EvaHan 2026 tasks, this study proposes an LVLM-based framework for printed/handwritten text recognition and layout analysis. To effectively adapt the Qwen2.5-VL-7B-Instruct model, our methodology innovates through a dual-level optimization strategy: distinct augmentation strategies are developed for OCR and layout tasks, while task-specific prompt templates are engineered to decouple text transcription from coordinate prediction. This combined approach significantly enhances overall task proficiency, achieving Character Error Rates of 0.0372 (printed) and 0.0823 (handwritten), alongside a mean average Precision of 0.2933 for layout analysis. Results show general LVLMs underperform in zero-shot ancient text tasks, but fine-tuning with tailored strategies significantly boosts performance and highlights their potential.

Keywords: Optical Character Recognition, Layout Analysis, Ancient Texts, Large Vision-Language Models

1. Introduction

Ancient Chinese texts are central to the study of Chinese civilization. Their large-scale digitization and intelligent use are vital for cultural preservation and academic research(Shi et al., 2025).However, processing these historical materials is highly challenging. Beyond the difficulties of recognizing variant characters and damaged strokes, deeply understanding the document layout is crucial for accurate digitization(Luo et al., 2024). This is particularly difficult in ancient texts due to their complex structural elements—such as the intermingling of text blocks, margins, seals, and illustrations—alongside mixed handwritten and printed scripts. Together, these factors pose significant challenges for traditional Optical Character Recognition (OCR) technologies.

In recent years, Large Vision-Language Models (LVLMs) have shown strong potential in document analysis and scene text recognition, thanks to their ability to jointly process visual and linguistic information. As a result, evaluating and improving LVLM performance on ancient document images has become a key research direction.

We present a systematic LVLM optimization framework for the EvaHan 2026 ancient Chinese OCR evaluation. By fine-tuning Qwen2.5-VL-7B-Instruct with LoRA and task-specific prompts , alongside tailored data augmentation—geometric transformations/hybrid stitching for text recognition and background embedding for layout analysis —our approach outperformed baselines across all tracks. Code is publicly available on GitHub[1].

2. Related Works

OCR research has undergone decades of iterative development. Early studies primarily relied on traditional machine learning algorithms, such as Support Vector Machines (SVM) and Hidden Markov Models (HMM). These methods achieved character recognition by extracting geometric and textural features. However, they exhibited poor adaptability to variant characters and degraded strokes in ancient texts, resulting in limited recognition accuracy. With the advancement of deep learning, OCR models based on Convolutional Neural Networks (CNN) and Transformers—such as CRNN (Shi et al., 2016) and TrOCR (Li et al., 2023) —have emerged as the predominant approaches in the field. By implementing end-to-end character detection and recognition, these models have made significant progress in printed text scenarios. Nevertheless, in complex scenarios involving handwritten scripts and mixed layouts, challenges such as disordered reading sequences and high character error rates persist.

Recently, the emergence of LVLMs has brought new breakthroughs to ancient text processing(Wang and Zhu, 2025). Several models, such as Qwen2.5-VL(Bai et al., 2025), simplify the workflow by directly processing "image-text" pairs through enhanced visual recognition and cross-modal fusion, bypassing the need for separate, multi-stage detection and recognition. Additionally, DocLayout-YOLO

[1]https://github.com/iamxiatian/evahan2026

Proceedings of LT4HALA 2026 @ LREC 2026, pages 299–304.
11 May 2026.

has substantially improved layout element analysis for general documents through data synthesis and model optimization(Zhao et al., 2024). However, most existing multimodal models are designed for general-purpose scenarios and lack specialized optimization for the unique element layouts and calligraphic styles of ancient texts. Consequently, their recognition precision and layout analysis capabilities in ancient document contexts remain to be validated. The EvaHan 2026 international evaluation has catalyzed applied research into LVLMs for ancient text processing, providing a unified benchmark for comparing the performance of diverse methodologies.

3. Methodology

3.1. Overall Framework and Baseline Model Selection

As illustrated in Figure 1, the proposed recognition and optimization framework for ancient text and layout elements consists of four interconnected stages: data augmentation, model fine-tuning, model inference, and result parsing. Within this pipeline, the data augmentation module processes raw images through geometric transformations, stitching, and background embedding to generate datasets more conducive to model learning. These augmented datasets are then fed into the model fine-tuning module, which utilizes the ms-swift framework and LoRA technology to derive optimized model weights. Subsequently, the inference module converts user inputs into model-compatible formats and generates textual outputs. Finally, the result parsing module analyzes the raw outputs from the LVLM, extracting key information to generate structured results.

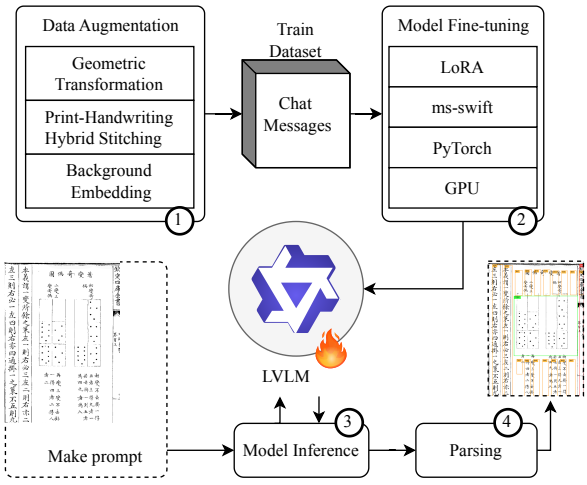

Figure 1: The overall architecture of the proposed recognition and optimization framework

In terms of baseline selection, we conducted a comparative analysis of the fine-tuning effectiveness between Qwen2.5-VL-7B-Instruct (Qwen) and Xunzi_Qwen2_VL_7B_Instruct (Xunzi) across the training sets of the three sub-tasks. Preliminary evaluations demonstrated that Qwen exhibited superior performance across all metrics; consequently, it was adopted as the primary backbone for this study. By integrating a native dynamic-resolution Vision Transformer and a window attention mechanism, this model effectively captures the multi-scale and high-dimensional complexity inherent in ancient Chinese document images.

3.2. Data Augmentation Strategies

Data augmentation serves as a critical mechanism for enhancing model generalization and performance. Given the specific characteristics of ancient Chinese texts, we implemented distinct augmentation strategies for the OCR tasks and the layout element analysis task.

3.2.1. Data Augmentation for OCR Tasks

For the closed track, we employed geometric transformation augmentation. Each image was subjected to a series of randomized operations, including horizontal/vertical flips, affine transformations, and random cropping. By generating 4 to 5 variants per original image, we expanded the training set to over 20,000 samples. This approach significantly bolsters the model's robustness to geometric distortions commonly found in scanned ancient documents.

For the open track, we introduced a hybrid print-handwriting stitching strategy. Pairs of images were sequentially selected from Dataset A (printed) and Dataset C (handwritten), denoted as a and c, respectively. These were then concatenated vertically in two ways (a above c, and c above a) to form new composite images. The width of the resulting image was determined by the larger of the two originals, and their corresponding text labels were concatenated accordingly. A total of 10,000 hybrid samples were generated. This strategy encourages the model to differentiate between the two calligraphic styles within a single input, effectively yielding an implicit multi-task learning effect.

3.2.2. Data Augmentation for Layout Element Analysis

For the closed track, we implemented a background embedding augmentation strategy to enhance spatial generalization and mitigate coordinate instability. This approach addresses two primary challenges. First, it prevents the model from memorizing absolute coordinates. Second, it mitigates stability issues caused by variations in input image

dimensions. Specifically, we initialize a random light-colored background canvas with fixed dimensions of 924 $\times$ 1232 pixels. Original images are stochastically positioned within this canvas; any image exceeding these dimensions is proportionally rescaled prior to placement. For each sample, two augmented variants are generated: one anchored at the top-left corner and another at a randomized location, ensuring the entire image remains within the canvas boundaries. By decoupling layout elements from fixed coordinate systems, this strategy forces the model to rely on intrinsic layout features, thereby improving localization robustness and spatial generalization.

For the open track, the training set was directly converted into the DocLayout-YOLO training format. We discarded the textual information contained within the layout elements of the original data and focused exclusively on generating bounding box coordinates and category labels for the layout elements in Dataset B.

3.3. Instruction Design and Fine-tuning Methodology

Instructions serve to guide the behavior of LVLMs and constrain their output space. In contrast to specialized traditional models, LVLMs function under a conversational generative paradigm, leveraging its inherent multi-task capabilities which are activated through specific prompt cues. Even after Parameter-Efficient Fine-Tuning (PEFT) such as LoRA, the model primarily optimizes the mapping between specific instructions and target task outputs, without altering its fundamental reliance on prompt-based guidance. While fine-tuning enables the model to acquire specialized competencies in text transcription and coordinate prediction, meticulously designed instructions are crucial for defining task roles, output formats, and specific constraints. Furthermore, in scenarios where a single model handles multiple sub-tasks, instructions act as the unique identifier for task differentiation, effectively suppressing hallucinatory outputs irrelevant to the evaluation criteria.

We designed differentiated prompt templates for the three sub-tasks to activate the model's fine-tuned knowledge and standardize the output formats. Specifically, for Task A and Task C, the prompts emphasize natural reading order, the handling of special characters, and the prohibition of extraneous content. For the Layout Analysis task, the instructions specify element categories, coordinate formats, and text content requirements, supplemented by examples to reduce the model's cognitive load. Detailed prompt templates are available in our open-source repository on GitHub. The model was fine-tuned using LoRA(Hu et al., 2022) , an PEFT technique that keeps the pre-trained weights frozen while introducing a minimal set of trainable parameters.Its core principle involves using low-rank factorization to simulate parameter updates during task adaptation. By updating only a fraction of the total parameters, LoRA significantly reduces GPU memory consumption and accelerates training efficiency. Both training and inference were implemented using the ms-swift framework(Zhao et al., 2025).

4. Experiments

4.1. Datasets

We strictly followed the official data partitioning of EvaHan 2026, utilizing the training set for fine-tuning and the test set for evaluation. The EvaHan 2026 dataset comprises three categories of image-text pairs: plain text images(Dataset A), mixed layout images(Dataset B), and handwritten text images(Dataset C). These high-quality datasets were generated through automated annotation followed by expert revision. For the OCR tasks, we merged Dataset A and Dataset C, along with their respective augmented variants, to form the training corpus. For the Layout Element Recognition task, Dataset B was first converted into a standardized labeling format to serve as training data for DocLayout-YOLO. Subsequently, these images were processed via background embedding to generate 10,000 samples for fine-tuning the LVLM.

4.2. Parameter Settings

Model fine-tuning was performed using the SWIFT framework, employing Parameter-Efficient Fine-Tuning via LoRA. The LoRA rank (r) was set to 8, with an alpha of 32, and the target modules encompassed all linear layers; this configuration represented a trade-off to optimize VRAM efficiency under experimental constraints. To determine the optimal status for the visual encoder, preliminary empirical evaluations were conducted during the initial two training epochs. Results indicated that unfreezing the ViT yielded no measurable performance gains while increasing computational overhead. Consequently, the visual encoder was frozen (freeze_vit=true) to optimize training efficiency and mitigate potential overfitting, restricting parameter updates to the language model and alignment layers. Key hyperparameters were standardized as follows: a learning rate of 2×10^{-5}, a global batch size of 32 (achieved via a per-device batch size of 2, gradient accumulation of 4, and 8 GPUs), 8 training epochs, the AdamW optimizer, and a warmup ratio of 0.05. Due to the variance in output lengths between OCR and layout tasks, the maximum sequence length was set to 2,048 for Tasks A/C and

8,192 for Task B. All experiments were conducted on a cluster of 8 NVIDIA A800 GPUs.

Regarding checkpoint selection, for the OCR tasks, we utilized the final checkpoint (Epoch 8) for inference. In contrast, for the layout analysis task, the model was trained for a total of 15 epochs with checkpoints saved every 100 steps. To mitigate overfitting and ensure optimal generalization, we selected an intermediate checkpoint at Step 1,900 (approximately the 12th epoch) for the final submission, as it demonstrated superior performance on a manually sampled validation set. During the inference phase, the vLLM toolkit(Kwon et al., 2023) was employed for acceleration, with the temperature set to 0.1, top_p to 0.9, and the maximum sequence length to 8,192. For the layout task, the model's raw HTML-tagged outputs underwent post-processing to extract bounding boxes (bboxes) and textual content, which were subsequently converted into the JSON format required for the official evaluation.

4.3. Evaluation Metrics

Following the official EvaHan 2026 evaluation metrics, we adopt CER, NED, F1-score, and C-Score for the OCR tasks (Tasks A and C). These metrics are calculated in two versions: considering and ignoring variant characters. For layout analysis (Task B), mAP@[.5:.95], Micro/Macro F1, and Average Match IoU are utilized. We refer readers to the official guidelines for detailed metric formulations[2].

4.4. Methods and Results

To evaluate the effectiveness of the proposed framework, we conducted comparative experiments involving both baselines and our proposed models. The detailed technical specifications of these eight methods are summarized in Table 1 in the Appendix.

As demonstrated in Tables 2, 3, and 4 (see Appendix), our proposed methods significantly outperform all baselines across all evaluated metrics. In the closed track, for Task A (including variant characters), the CER decreased from 0.0618 (baseline) to 0.0372—a substantial reduction of 39.8%—while the F1-score rose to 0.9663, marking a 2.47% improvement over the baseline's 0.9430. For Task C (including variants), the CER dropped from 0.0920 to 0.0823 (a 10.5% decrease), with the F1-score increasing by 1.08% to reach 0.9197. Most notably, in Task B, the proposed method demonstrated significant improvements in mAP, Micro F1, Macro F1, and Avg Match IoU, with improvements of 46.21%, 370.57%, 145.23%, and 7.30%, respectively, compared to the best-performing baseline. These results validate the efficacy of our geometric transformation and background embedding strategies.

In the open track results, the proposed hybrid print-handwriting stitching strategy consistently surpassed the best baseline results for both Task A and Task C, although its performance was slightly lower than that of the geometric transformation augmentation. For Task B, we fine-tuned DocLayout-YOLO, a specialized layout analysis model pre-trained on 300K synthetic document pages. This model employs a Controllable Receptive Field module and a Global-to-Local design to refine multi-scale features. After fine-tuning on Dataset B, it achieved a mAP@[.5:.95] of 0.3828 and Micro F1 of 0.6673, substantially outperforming the closed-track LVLM (0.2933 mAP and 0.2414 Micro F1). This demonstrates the advantage of specialized models for precise layout analysis.

4.5. Discussion

LVLMs provide a powerful means of understanding the layout composition and textual content of ancient document images. For OCR tasks, LVLMs demonstrate robust baseline recognition capabilities due to extensive pre-training, which can be further refined through fine-tuning. However, in layout element recognition, the zero-shot performance of LVLMs is relatively poor; while fine-tuning yields substantial improvements over the baseline, the overall performance remains suboptimal.

Our experiments reveal that meticulous instruction design significantly reduces output errors, indicating that LVLMs are highly sensitive to prompts. Consequently, task design should strategically leverage the model's in-context learning capabilities. Furthermore, targeted fine-tuning on domain-specific data consistently enhances performance, though the efficacy is heavily influenced by specific data augmentation techniques and instruction strategies. In layout analysis, the background embedding strategy, which enforces the learning of relative spatial layouts, resulted in a substantial improvement in performance compared to the baseline. For OCR tasks—encompassing both printed and handwritten scripts—our geometric transformation and hybrid stitching strategies outperformed the baseline models. Notably, the hybrid stitching strategy was slightly less effective than geometric transformation, likely because direct stitching disrupts textual coherence; future research could explore more natural synthesis methods for mixed-media data.

In the context of layout analysis, LVLMs currently face higher training costs and slower inference speeds, with precision currently remaining below that of specialized layout-specific architectures.

[2] https://github.com/GoThereGit/EvaHan

Nevertheless, their primary advantage lies in end-to-end semantic understanding. Future work could focus on constructing hybrid architectures that combine the strengths of both specialized detectors and multimodal reasoners.

Regarding variant characters, the marginal performance gap between "variant-aware" and "standard" metrics in printed texts suggests that the model recognizes printed characters with high accuracy and is minimally affected by variant interference. In contrast, for handwritten scripts, this gap exceeds one percentage point. This indicates that in over 1% of cases, the model's output is semantically appropriate but fails to map to the specific character in the variant set. This ambiguity aligns with the inherent variability in calligraphic styles across different scribes, presenting challenges analogous to those encountered in human expert decipherment of ancient manuscripts. Overall, the model performs better under variant-aware evaluation conditions, demonstrating strong morphological generalization and an ability to correctly identify and map variants to their standard counterparts.

5. Conclusion and Future Work

This paper presents a robust framework based on LVLMs, optimized through strategic data augmentation and instruction fine-tuning. Our methods achieved performance that significantly surpassed the baseline models across all three tasks of the EvaHan 2026 international evaluation, thereby validating the efficacy of the proposed approach. Experimental evidence highlights that tailored augmentation strategies, such as background embedding and hybrid stitching, are crucial for adapting general-purpose LVLMs to the nuances of ancient Chinese texts.

Furthermore, as the test set ground truth is currently withheld by the evaluation organizers, we are temporarily unable to present comprehensive ablation and error analyses. In the future, we plan to conduct systematic ablation experiments to quantify the individual contributions of each proposed module, alongside an in-depth analysis of failure cases to guide further optimizations.

In future work, we also intend to explore the fusion of diverse data augmentation strategies and investigate more sophisticated techniques to address the remaining challenges in spatial localization for layout elements within large models. By refining these mechanisms, we aim to propel the digitization of ancient documents toward higher precision and broader academic utility.

6. Acknowledgements

This research is supported by the National Social Science Fund of China (22BTQ068).

7. Bibliographical References

Shuai Bai, Keqin Chen, Xuejing Liu, et al. 2025. Qwen2.5-VL technical report.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, et al. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Minghao Li, Tengchao Lv, Jingye Chen, et al. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 13094–13102.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, et al. 2024. LayoutLLM: layout instruction tuning with large language models for document understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15630–15640.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.

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

Dongbo Wang and Dongmei Zhu. 2025. Benchmarking the ancient books capability of multimodal large language models. *Npj Heritage Science*, 13(1):339.

Yuze Zhao, Jintao Huang, Jinghan Hu, et al. 2025. SWIFT: A scalable lightweight infrastructure for fine-tuning.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, et al. 2024. DocLayout-YOLO: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception.

A. Appendix: Detailed Experimental Results

Method	**Backbone**	**FT Method**	**Technical Strategy**	**Target Task**
Baseline-A	Qwen2.5-VL-7B	Zero-shot	Zero-shot instruction engineering without fine-tuning.	Official Baseline
Baseline-B	Qwen2.5-VL-7B	LoRA	Standard LoRA fine-tuning on task-specific datasets.	Official Baseline
Baseline-C	Xunzi_Qwen2_VL_7B	Zero-shot	Zero-shot instruction engineering without fine-tuning.	Official Baseline
Baseline-D	Xunzi_Qwen2_VL_7B	LoRA	Standard LoRA fine-tuning on task-specific datasets.	Official Baseline
ACBM_GT	Qwen2.5-VL-7B	LoRA	Incorporate Geometric Transformation (GT) augmentation.	Closed: Task A & C
ACBM_HS	Qwen2.5-VL-7B	LoRA	Utilize Hybrid Stitching (HS) for print and handwritten scripts.	Open: Task A & C
ACBM_BE	Qwen2.5-VL-7B	LoRA	Employ Background Embedding (BE) by integrating document samples into light-colored backgrounds.	Closed: Task B
ACBM_YOLO	DocLayout-YOLO	Full Fine-tuning	Employ DocLayout-YOLO-DocLayNet-Docsynth300K_pretrained model for layout element detection.	Open: Task B

Table 1: Description of Comparative Methods

Method	Consider variant characters				Don't consider variant characters			
	CER	NED	F1	C-Score	CER	NED	F1	C-Score
Baseline-A	0.1014	0.0947	0.9110	0.9037	0.1121	0.1054	0.9007	0.8931
Baseline-B	0.0618	0.0613	0.9430	0.9397	0.0685	0.0679	0.9364	0.9331
Baseline-C	0.1786	0.1740	0.8409	0.8282	0.1851	0.1802	0.8345	0.8218
Baseline-D	0.1214	0.1183	0.8993	0.8854	0.1264	0.1232	0.8945	0.8805
ACBM_GT(Close)	**0.0372**	**0.0369**	**0.9663**	**0.9639**	**0.0382**	**0.0379**	**0.9653**	**0.9629**
ACBM_HS(Open)	**0.0476**	**0.0467**	**0.9560**	**0.9537**	**0.0490**	**0.0481**	**0.9546**	**0.9523**

Table 2: Results of Task A

Method	mAP@[.5:.95]	Micro F1	Macro F1	Avg Match IoU
Baseline-A	0	0	0	0
Baseline-B	0.2006	0.0513	0.1530	0.6600
Baseline-C	0.0236	0.0003	0.0114	0.5943
Baseline-D	0.1917	0.0403	0.1130	0.6654
ACBM_BE(Close)	**0.2933**	**0.2414**	**0.3752**	**0.7140**
ACBM_YOLO(Open)	**0.3828**	**0.6673**	**0.5983**	**0.8041**

Table 3: Results of Task B

Method	Consider variant characters				Don't consider variant characters			
	CER	NED	F1	C-Score	CER	NED	F1	C-Score
Baseline-A	0.1207	0.1193	0.8849	0.8812	0.1338	0.1324	0.8718	0.8681
Baseline-B	0.0920	0.0919	0.9099	0.9086	0.1066	0.1065	0.8953	0.8940
Baseline-C	0.1497	0.1492	0.8538	0.8514	0.1609	0.1604	0.8425	0.8402
Baseline-D	0.1383	0.1376	0.8673	0.8635	0.1520	0.1512	0.8534	0.8498
ACBM_GT(Close)	**0.0823**	**0.0822**	**0.9197**	**0.9183**	**0.0952**	**0.0951**	**0.9068**	**0.9054**
ACBM_HS(Open)	**0.0853**	**0.0849**	**0.9168**	**0.9154**	**0.1030**	**0.1026**	**0.8991**	**0.8977**

Table 4: Results of Task C