A Multi-Stage System for Ancient Chinese OCR and Layout Understanding in the EvaHan2026 Shared Task

KeYan Liang, Meiling Liu
Aulin College, Northeast Forestry University
Harbin, Heilongjiang, China
eyonryan00@nefu.edu.cn, mlliu@nefu.edu.cn

Abstract

This paper presents a multi-stage system for the EvaHan2026 shared task, addressing the complex challenges of ancient Chinese optical character recognition (OCR) and layout understanding. For text recognition (Tasks A and C), we adopt parameter-efficient LoRA fine-tuning on the Qwen2.5-VL-7B-Instruct vision-language model (VLM). By directly processing full-resolution long-column images, we preserve critical spatial and contextual integrity without heuristic region cropping. For document layout analysis (Task B), we propose a novel hybrid perception-reasoning paradigm. Instead of relying solely on scaling visual detectors, we decouple localization and understanding: utilizing a YOLO-based ensemble for precise spatial bounding, and casting the VLM as a semantic verifier to eliminate spurious detections. Evaluated on the official unseen test set, our system achieves substantial improvements over the provided baselines, obtaining a 0.0441 Character Error Rate (CER) for printed OCR, a 0.0793 CER for handwritten OCR (including variants), and a 0.5118 mAP@[0.5:0.95] for layout detection. These results demonstrate that integrating VLM-based semantic reasoning into traditional visual detection pipelines is highly effective for multimodal historical document analysis.

Keywords: Ancient Chinese OCR, Layout Analysis, Vision-Language Model, Hybrid Perception-Reasoning, EvaHan2026

1. Introduction

Digitizing ancient Chinese documents poses substantial computational challenges due to complex layouts, severe degradation, cursive stroke adhesion, and variant glyphs. Furthermore, diverse visual elements (text, seals, illustrations) frequently occlude one another, complicating document understanding. The EvaHan2026 shared task provides a benchmark for this. Tasks A and C focus on printed and handwritten OCR within pre-cropped vertical columns (often reaching extreme aspect ratios of approximately 200×4000 pixels), requiring models to preserve long-context spatial information. Conversely, Task B requires full-page layout detection across four distinct categories.

In this paper, we propose a robust multi-stage system tackling these subtasks. Beyond pure engineering optimizations, our core methodological insight is decoupling spatial localization from semantic understanding. Our main contributions are:

- **Direct Long-Column OCR via VLM Adaptation:** For Tasks A and C, we adapt a state-of-the-art vision-language model (VLM), Qwen2.5-VL (Bai et al., 2025), to the ancient document domain via parameter-efficient LoRA fine-tuning. By treating single-column OCR as a direct image-to-sequence generation task on full-resolution images, we entirely bypass error-prone character-level cropping pipelines.
- **Hybrid Perception-Reasoning Layout Paradigm:** For Task B, we introduce a category-specific hybrid framework casting the VLM as a semantic critic. While a YOLO-based dual-stream ensemble (Ultralytics, 2023, 2024) handles raw spatial localization, the VLM acts as a semantic verifier to filter out hard false positives. Combined with a two-stage seal classification pipeline and Hough-based geometric reasoning, this paradigm effectively bridges raw object detection and high-level cognitive verification.

2. Related Work

Traditional cascaded OCR pipelines often suffer from error propagation. Recently, Large Vision-Language Models (VLMs) (Bai et al., 2023; Wang et al., 2024; Bai et al., 2025) have shown remarkable multimodal capabilities. We build upon this by applying LoRA-adapted VLMs (Hu et al., 2022) directly to uncropped, full-resolution ancient text columns. For Document Layout Analysis (DLA), while unified toolkits (Shen et al., 2021) and multimodal frameworks (Xu et al., 2020) advance modern document understanding, historical scans still heavily rely on object detectors (Ultralytics, 2023, 2024) and text detectors (Liao et al., 2020) for bounding box precision. We overcome the limitations of pure visual detection and end-to-end models (Li et al., 2021; Blecher et al., 2023) in dense scans by introducing a VLM-in-the-loop verification step, bridging raw detection and semantic reasoning. Finally, we mitigate VLM memory bottlenecks using vLLM (Kwon et al., 2023) and address se-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 263–267.

vere class imbalances via targeted synthetic data augmentation (Shorten and Khoshgoftaar, 2019).

3. Methodology

3.1. Full-Resolution OCR Adaptation (Tasks A and C)

For both printed text OCR (Task A) and handwritten calligraphy OCR (Task C), the provided document images are pre-cropped into single vertical text columns. While this isolates the text sequence, the extreme aspect ratios (e.g., approximately 200×4000 pixels) challenge traditional recognition pipelines. We directly fine-tune Qwen2.5-VL-7B-Instruct (Bai et al., 2025) using parameter-efficient Low-Rank Adaptation (LoRA). The base model parameters remain completely frozen, and only the low-rank adaptation matrices are updated. We configure the LoRA hyperparameters with a rank $r = 64$, a scaling factor $\alpha = 128$, and a dropout rate of 0.05. This strategy mitigates catastrophic forgetting, minimizes trainable parameters, and demonstrates robust domain adaptation.

3.2. Multi-Stage Layout Detection (Task B)

Task B focuses on detecting four distinct layout categories. Unlike conventional parallel detection pipelines, our system implements a strictly sequential, three-phase execution flow (Figure 1). This design explicitly models the strong spatial dependencies between different layout elements.

3.2.1. Phase 1: Primary Text Detection

The *Text Pipeline* executes first, providing essential spatial priors for downstream modules. Initially, the RGB image and its adaptively binarized version are fed into a dual-stream YOLO ensemble (YOLOv8 and YOLO11). To mitigate high false-positive rates in historical text detection, Qwen2.5-VL acts as a semantic verifier, evaluating each candidate box to confirm visible Chinese characters and reject artifacts. Finally, Non-Maximum Suppression (NMS) and morphological cleaning merge erroneously split components.

3.2.2. Phase 2: Parallel Multimodal Pipelines

Conditioned on Phase 1 outputs, the remaining pipelines execute in parallel:

Image Pipeline. Following YOLO ensemble detection, we enforce two text-dependent geometric heuristics: (1) *Overlap Rejection*: Discarding image boxes with $> 80\%$ IoU with text regions; (2) *Density Rejection*: Removing candidates encompassing ≥ 4 text boxes that cover $> 45\%$ of the image area. Notably, our VLM verifier is exclusively applied to the *text* category to optimally balance reasoning capabilities with system latency, as other categories are better suited for geometric or lightweight classification approaches.

Seal Pipeline. We employ a dedicated two-stage cascade to balance recall and precision for degraded seals. Following a high-recall YOLO detection and NMS, each candidate crop is adaptively binarized, padded to a standardized 256×256 square, and fed into a specialized YOLO classification model (confidence threshold > 0.80).

Book Edge Pipeline. After initial YOLO detection, we convert the region to grayscale, remove black borders, and apply the Hough Transform to extract candidate vertical spine lines, followed by a lateral whiteness density verification.

3.2.3. Phase 3: Global Conflict Resolution

In the final phase, a pipeline merger standardizes the heterogeneous outputs and resolves spatial collisions across all categories using a deterministic cross-class rule set:

1. **Seal overrides Image:** When a seal and an image overlap significantly (IoU > 0.65), the image bounding box is considered a false positive and deleted.
2. **Text Excludes Seal:** To prevent dense text blocks from being misclassified as false seals, any seal box whose 90% area encompasses two or more text boxes is deleted.
3. **Book Edge Physical Cropping:** Any layout object overlapping with a book edge by more than 70% is entirely removed. For partial occlusions ($\leq 70\%$), the object's bounding box is geometrically cropped along the spine line.

4. Experiments

4.1. Dataset and Augmentation

The EvaHan2026 dataset consists of 5,000 samples for each subtask, partitioned into 4,500 for training and 500 for validation. Tasks A and C represent samples as single vertical columns (200×4000 pixels), while Task B provides binarized full-page scans (500×900 pixels).

To mitigate the severe class imbalance in the *seal* category (only $\sim$600 labels), we implemented a targeted synthetic data generation pipeline. By applying a $5\times$ augmentation factor, we created 3,000 layout images. We simulated historical degradation

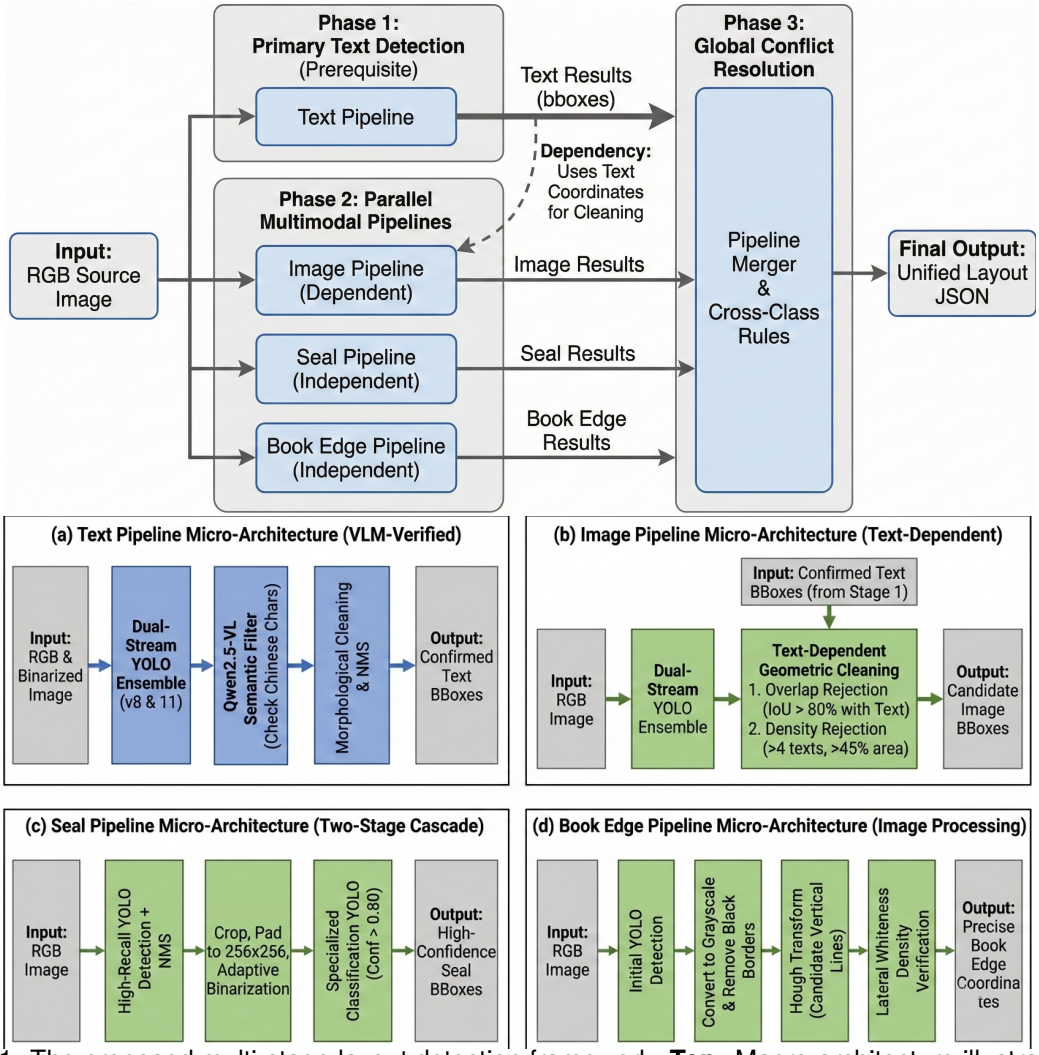

Figure 1: The proposed multi-stage layout detection framework. **Top:** Macro-architecture illustrating the sequential execution flow. **Bottom:** Micro-architectures of the category-specific pipelines.

through adaptive Gaussian thresholding and morphological erosion to mimic ink fading, blending the synthetic seals into background documents. While this introduces domain gaps, our high-recall initial detection coupled with the secondary classification network effectively filters out spurious detections.

4.2. Implementation Details

Experiments were conducted on a single NVIDIA RTX 5090 GPU (32GB VRAM). For Tasks A and C, we fine-tuned Qwen2.5-VL using LLaMA-Factory (Zheng et al., 2024) with FlashAttention-2 (Dao, 2023). The model was trained in bf16 precision (batch size 1, 8 gradient accumulation steps, 5×10^{-5} peak learning rate). Task C required 2,400 training steps, whereas Task A converged at 1,400 steps. For Task B, a YOLOv8-L and a YOLOv11-X model were trained for 600 epochs. To manage the severe memory bottlenecks of deploying a 7B VLM concurrently with YOLO models, we utilized `vLLM`, restricting its GPU memory utilization to 70%.

4.3. Results

Our system demonstrates robust performance, substantially outperforming official baselines. Rigorous empirical observations on our local validation split guided the design of our category-specific pipelines.

Task A and Task C (OCR). Table 1 compares our fine-tuning strategy against official baselines. For printed text (Task A), our system establishes a new state-of-the-art (CER 0.0441), significantly reducing errors. To address the performance gap between Task A and Task C, we conducted a qualitative error analysis on the handwritten calligraphy samples. We observed that errors are predominantly driven by morphological challenges rather than semantic ones. The extreme cursive fluidity and severe stroke adhesion in historical calligraphy present significant visual ambiguities, heavily disrupting the spatial recognition sequence.

Subtask	System	Including Variants				Excluding Variants			
		CER ↓	NED ↓	F1 ↑	Score ↑	CER ↓	NED ↓	F1 ↑	Score ↑
Task A	Qwen2.5-VL-7B (Zero-shot)	0.1014	0.0947	0.9110	0.9037	0.1121	0.1054	0.9007	0.8931
	Qwen2.5-VL-7B (Instr. Tuning)	0.0618	0.0613	0.9430	0.9397	0.0685	0.0679	0.9364	0.9331
	Xunzi-Qwen2-VL-7B (Zero-shot)	0.1786	0.1740	0.8409	0.8282	0.1851	0.1802	0.8345	0.8218
	Xunzi-Qwen2-VL-7B (Instr. Tuning)	0.1214	0.1183	0.8993	0.8854	0.1264	0.1232	0.8945	0.8805
	Ours (LoRA Fine-Tuned)	**0.0441**	**0.0435**	**0.9589**	**0.9569**	**0.0451**	**0.0446**	**0.9579**	**0.9559**
Task C	Qwen2.5-VL-7B (Zero-shot)	0.1207	0.1193	0.8849	0.8812	0.1338	0.1324	0.8718	0.8681
	Qwen2.5-VL-7B (Instr. Tuning)	0.0920	0.0919	0.9099	0.9086	0.1066	0.1065	0.8953	0.8940
	Xunzi-Qwen2-VL-7B (Zero-shot)	0.1497	0.1492	0.8538	0.8514	0.1609	0.1604	0.8425	0.8402
	Xunzi-Qwen2-VL-7B (Instr. Tuning)	0.1383	0.1376	0.8673	0.8635	0.1520	0.1512	0.8534	0.8498
	Ours (LoRA Fine-Tuned)	**0.0793**	**0.0791**	**0.9222**	**0.9212**	**0.0917**	**0.0915**	**0.9098**	**0.9088**

Table 1: Comprehensive evaluation results for OCR subtasks (Task A and Task C) on the official test set. Our parameter-efficient fine-tuning strategy consistently outperforms all official baselines across both variant settings.

System	Method	mAP@[.5:.95] ↑	Micro F1 ↑	Macro F1 ↑	Avg Match IoU ↑
Qwen2.5-VL-7B	Zero-shot	0.0000	0.0000	0.0000	0.0000
Qwen2.5-VL-7B	Instr. Tuning	0.2006	0.0513	0.1530	0.6600
Xunzi-Qwen2-VL-7B	Zero-shot	0.0236	0.0003	0.0114	0.5943
Xunzi-Qwen2-VL-7B	Instr. Tuning	0.1917	0.0403	0.1130	0.6654
Ours	**Hybrid Multi-Stage Pipeline**	**0.5118**	**0.7553**	**0.7879**	**0.7740**

Table 2: Performance metrics for Layout Detection (Task B). Our hybrid pipeline, which integrates YOLO ensemble detection with VLM-based semantic verification, substantially outperforms all official baselines.

Task B (Layout Detection). As shown in Table 2, our hybrid pipeline yields an mAP@[0.5:0.95] of 0.5118. Compared to the best-performing baseline (0.2006 mAP), our system achieves a relative improvement of over 155%. This substantial improvement validates the efficacy of our category-aware geometric cleaning and conflict resolution module.

5. Conclusion

In this paper, we presented a highly effective, multi-stage system for the EvaHan2026 shared task. For text recognition (Tasks A and C), applying parameter-efficient LoRA fine-tuning to Qwen2.5-VL on full-resolution images successfully preserves crucial spatial context, achieving highly competitive CERs. For layout analysis (Task B), we developed a category-aware hybrid framework integrating a YOLO ensemble, a specialized seal pipeline, and a VLM-based semantic verifier. This approach successfully mitigated the high false-positive rates typical of historical documents, highlighting the immense potential of combining robust visual detection with VLM semantic reasoning.

6. Limitations and Future Work

Several limitations remain for future exploration. First, while our models train on pre-cropped columns (Tasks A/C), future work will target end-to-end full-page OCR to improve generalization in uncropped scenarios. Second, our deterministic spatial conflict rules are highly optimized for the EvaHan2026 dataset; learning-based spatial relation modules could enhance zero-shot generalizability. Third, expanding the VLM semantic verifier beyond the *text* category to caption or verify other layout elements remains a promising direction. Finally, addressing label noise and generating more diverse synthetic data for stylistically homogeneous categories like seals will further bridge domain gaps.

7. Acknowledgements

We gratefully acknowledge the EvaHan2026 organizers for providing the standardized benchmark and dataset.

8. Ethics Statement

This research aids the automated digitization of historical Chinese literature. All datasets utilized were officially provided by the EvaHan2026 organizers and consist entirely of ancient, public-domain texts containing no sensitive information. The deployment of this system poses no negative societal impacts.

9. Bibliographical References

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 understand-

ing, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

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*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 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, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11474–11481.

Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. Layoutparser: A unified toolkit for deep learning based document image analysis. In *Document Analysis and Recognition – ICDAR 2021*, pages 131–146. Springer.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60. Article 60.

Ultralytics. 2023. Ultralytics yolov8 docs. https://docs.ultralytics.com/models/yolov8/. Official documentation.

Ultralytics. 2024. Ultralytics yolo11 docs. https://docs.ultralytics.com/models/yolo11/. Official documentation.

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*.

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 and Data Mining*, pages 1192–1200.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410. Association for Computational Linguistics.