A Parameter-Efficient and Data-Centric Framework for Ancient Chinese Text Recognition and Layout Analysis

Yuchun Meng
School of Information Management, Nanjing University, China
163 Xianlin Road, Qixia District, Nanjing, Jiangsu Province, 210023
mengyuchun@smail.nju.edu.cn

Abstract

This paper presents the system developed for the EvaHan 2026 shared task on Ancient Chinese OCR and Layout Analysis. Participating in the Closed Track, we propose a highly parameter-efficient, data-centric framework based on the Qwen2.5-VL-7B-Instruct multimodal large language model (MLLM). While the official baseline utilizes the same backbone architecture, our approach significantly outperforms it by integrating orientation-aware image preprocessing and expert-constrained adaptive prompt engineering. We employed Low-Rank Adaptation (LoRA) with a minimal rank configuration (Rank=16) to train three independent, task-specific adapters. Our system achieved exceptional results, recording an Overall score of 0.9703 and an F1-score of 97.19% on printed text recognition (Task A)—effectively halving the baseline's Character Error Rate. On handwritten texts (Task C), we maintained a highly competitive 90.18% F1-score. Furthermore, our model achieved significant progress in layout analysis (Task B), surpassing the baseline's Macro F1 by 172% (0.4162 vs. 0.1530) and mAP by 37%. These results underscore that embedding explicit document structure and semantic constraints into MLLMs is more effective than simply scaling model parameters.

Keywords: EvaHan 2026, Data-Centric AI, Ancient Chinese OCR, Multimodal Large Language Model, LoRA Fine-tuning, Prompt Engineering

1. Introduction

Ancient Chinese documents, such as the *Siku Quanshu* and Buddhist scriptures, are invaluable cultural heritage. However, automating their digitization is challenging due to complex layouts, diverse font styles (e.g., woodblock printing vs. handwriting), and the prevalence of variant characters. Building upon previous evaluations of ancient Chinese processing (Li et al., 2022), the EvaHan 2026 shared task addresses these challenges through three subtasks: Dataset A (Printed Characters), Dataset B (Layout Analysis), and Dataset C (Handwritten Characters).

In this work, we propose a unified end-to-end approach based on Qwen2.5-VL, a Vision-Language Model capable of processing high-resolution images and generating structured text responses. Unlike traditional OCR pipelines that rely on separate detection and recognition modules (e.g., CNN+RNN+CTC or encoder-decoder architectures) (Li et al., 2021), our method treats OCR and layout analysis as multimodal generation tasks. By applying instruction tuning with LoRA (Hu et al., 2021), we effectively adapted the general-purpose MLLM within the strict constraints of the Closed Track, aligning with recent advancements in OCR-free document understanding paradigms (Liu et al., 2024).

2. Methodology

2.1 Model Architecture

We selected Qwen2.5-VL-7B-Instruct (Wang et al., 2024) as our base model. This model features a sophisticated Vision Transformer (ViT) encoder utilizing Naive Dynamic Resolution, which handles variable-resolution images without distortion. This makes it particularly suitable for ancient books with varying aspect ratios. The visual features are projected into the language model's embedding space, allowing the MLLM to generate text or coordinate descriptions directly from visual inputs.

2.2 Data-Driven Preprocessing and Adaptive Prompt Engineering

Instead of applying a generic preprocessing pipeline, we adopted a data-driven strategy to tailor our model inputs and prompts. Prior to fine-tuning, we developed a diagnostic toolkit to profile the geometric characteristics of the dataset (e.g., aspect ratios and orientation distributions).

Statistical Analysis & Orientation Correction:

Our diagnostic script revealed a significant divergence in image layouts between datasets.

For Task C (Handwritten Slips): The analysis showed an extreme average aspect ratio (Width/Height$\approx$0.12), confirming that the majority of images were "tall and narrow" vertical slips. However, a subset of images ($N_{horizontal}$) were identified as horizontally aligned, which contradicted the standard vertical reading order of the pre-trained model. To address this divergence and accommodate the vertical reading bias of the pre-trained model, we applied an automated batch rotation pipeline to all training and testing images. This preprocessing step successfully unified all inputs into a consistent vertical orientation, ensuring strict geometric alignment between the visual features and the model's pre-training objective.

For Task A (Printed Texts): The diagnostic report indicated a more standard aspect ratio (0.73), guiding us to use a resizing strategy distinct from Task C.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 294–298.

Precision Prompt Engineering:

Leveraging the insights from our data profiling, we crafted structure-aware prompts that explicitly informed the model about the visual nature of the input.

Generic Prompt: *"OCR this image." (Performance: Low)*

Our Adaptive Prompt (Task C): *"You are an expert in digital humanities. The input image is a high-resolution vertical slip (aspect ratio ≈ 0.12) containing 1-50 handwritten Traditional Chinese characters. Please strictly preserve variant characters (e.g., '無', '為') and output the text in top-to-bottom order."*

To ensure inference efficiency and flexibility, these geometric priors (e.g., aspect ratio≈0.12, 1-50 characters) are not dynamically computed per image during inference. Instead, they act as static constants derived from the global dataset averages.

By explicitly embedding the statistical priors obtained from our code into the prompt, we significantly reduced the model's hallucination rate and improved the alignment between visual features and textual generation, aligning with recent findings on the efficacy of structure-aware prompt engineering in multimodal tasks (Awadalla et al., 2023).

2.3 Fine-tuning Strategy

The overall architecture of our data-centric, parameter-efficient pipeline is illustrated in Figure 1 below.

To prevent catastrophic forgetting and ensure task-specific optimization, we employed Low-Rank Adaptation (LoRA). We trained three separate, highly parameter-efficient adapters:

Adapter A: Fine-tuned on Dataset A for printed text recognition.

Adapter B: Fine-tuned on Dataset B for layout detection.

Adapter C: Fine-tuned on Dataset C for handwritten text recognition.

We strictly adhered to the Closed Track rules, using no external data or pre-trained OCR models other than the permitted base model.

3. Experiments

3.1 Experimental Settings

Platform: Alibaba Cloud PAI-DSW

GPU: NVIDIA A10 (24GB)
Base Model: Qwen2.5-VL-7B-Instruct
LoRA Config: Rank = 16, Alpha = 32, Dropout = 0.05
Target Modules: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Learning Rate: 2e-4 (with cosine decay)
Batch Size: 1 (with Gradient Accumulation Steps = 16)
Max Pixels:~1,200,000 (approx.)
Epochs: 3
Given the extremely high-resolution nature of ancient document images (Max Pixels ≈ 1,200,000) and the hardware constraint of a single 24GB VRAM GPU (NVIDIA A10), full-

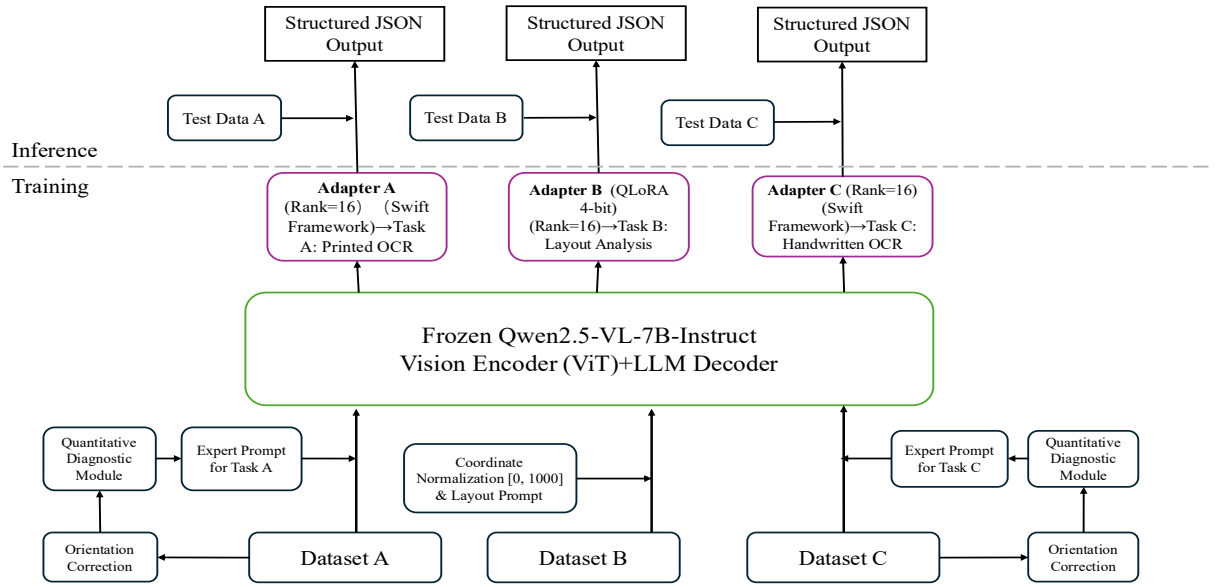

Figure 1 : *The unified multimodal framework. The pipeline is divided into training and inference phases. Raw datasets undergo quantitative diagnostics and task-specific geometric normalization (e.g., coordinate scaling for Task B). We explicitly apply parameter-efficient fine-tuning via three distinct LoRA adapters (Rank=16) to a frozen Qwen2.5-VL-7B backbone, utilizing QLoRA 4-bit precision and the Swift framework to manage hardware constraints.*

precision fine-tuning of a 7B multimodal model was computationally prohibitive. For Task A and C, we utilized the highly optimized Swift LLM framework (Qiu et al., 2023) to manage memory efficiently during training. Crucially, for the computationally intensive layout analysis task (Task B), we implemented 4-bit Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023). This mixed-precision engineering effectively compressed the frozen base model's memory footprint, allowing us to maintain high-resolution visual encoding without encountering Out-Of-Memory (OOM) errors, ensuring stable convergence over the 8-hour training cycle.

3.2 Results

We evaluated our system against the official baseline provided by the organizers, which also utilizes Qwen2.5-VL-7B-Instruct via the Swift framework but employs a heavier LoRA Rank of 32 on all-linear modules without specific data orientation preprocessing.

Task	Metric	Official Baseline (Rank 32)	Ours (Rank 16)	Improvement
Task A (Print)	CER ↓	0.0618	**0.0305**	-50.6%
	F1-Score ↑	0.9430	**0.9719**	+3.06 %
	Overall ↑	0.9397	**0.9703**	+3.25 %
Task B (Layout)	mAP ↑	0.2006	**0.2749**	+37.0 %
	Macro F1 ↑	0.1530	**0.4162**	+172.0 %
	Avg Match IoU ↑	0.6600	**0.6740**	+2.12 %
Task C (Hand)	CER ↓	**0.0920**	0.1010	-
	F1-Score ↑	**0.9099**	0.9018	-
	Overall ↑	**0.9086**	0.9000	-

Table 1 : Performance Comparison (Closed Track)(%)

As shown in Table 1, our system achieved strong performance. Most notably on Task A, our method halved the Character Error Rate (CER) to 3.05% and achieved an Overall score of 0.9703. On the highly challenging generative layout analysis task (Task B), we substantially outperformed the official baseline, improving the Macro F1 score by a substantial 172% (0.4162 vs. 0.1530). For handwritten texts (Task C), our model maintained highly competitive parity with the baseline (>90% F1), despite utilizing significantly fewer trainable parameters.

4. Discussion

4.1 Parameter Efficiency and Data-Centric Superiority

A critical insight from our experiments is the triumph of data-centric preprocessing over raw parameter scaling. The official baseline employed a higher LoRA Rank (32) and a large global batch size of 32. In contrast, our approach utilized a lightweight setup with a LoRA Rank of 16 targeting specific projection layers, coupled with a smaller effective batch size (Batch Size 1 with 16 accumulation steps). Despite using considerably fewer trainable parameters, our model significantly outperformed the baseline on Tasks A and B. This observation supports the premise that lower intrinsic rank and smaller batch sizes can sometimes act as a regularizer, preventing overfitting on specialized, low-resource datasets like ancient texts (Masters and Luschi, 2018; Liu et al., 2024). We attribute this success to the synergy between our orientation-aware image preprocessing and expert-constrained prompt engineering. We executed a batch rotation to unify the physical geometry of horizontal texts with the model's vertical reading priors. Furthermore, by reinforcing this with strict top-to-bottom layout constraints in the prompt, we effectively guided the MLLM's visual attention. This demonstrates that aligning the physical spatial structure of the input with the semantic constraints of the prompt is crucial for unlocking the latent capabilities. This phenomenon is corroborated by recent studies on visual grounding through structured prompting (Yang et al., 2024), and far surpasses the baseline's agnostic approach.

The Semantic Correction Phenomenon: Furthermore, the official evaluation data provides compelling evidence for the MLLM's intrinsic "semantic correction" capability. We observed a distinct performance gap in Task C (handwritten texts) depending on the evaluation metric: our system scored significantly higher when including variant characters (F1: 0.9018, Overall: 0.9000) compared to excluding them (F1: 0.8873, Overall: 0.8854). Interestingly, this gap was entirely absent in Task A, where the highly standardized woodblock fonts left little room for visual ambiguity. This contrast confirms that our lightweight adapter did not merely memorize isolated pixel shapes. Instead, when confronting the severe visual ambiguity inherent in handwritten historical manuscripts, the model successfully mapped the visual features of variant characters to their

normative semantic equivalents in the MLLM's latent space—a valuable capability for digital philology.

4.2 The Inherent Limits of Generative Coordinate Regression

On Task B, our model demonstrated exceptional categorization ability, outperforming the baseline's Macro F1 by 172% (0.4162 vs. 0.1530). However, the absolute mAP score remains modest at 0.2749. This reveals a fundamental characteristic of current Vision-Language Models: while MLLMs excel at semantic classification and rough regional identification (evidenced by our high F1 and 0.6740 Avg Match IoU), treating continuous spatial coordinates as discrete text tokens (Chen et al., 2021) leads to suboptimal pixel-level bounding box regression. In our architecture, the absolute pixel coordinates are mapped into 1000 discrete bins (i.e., coordinate normalization [0, 1000]). This discretization, inherent to MLLMs, inherently limits pixel-perfect localization. Furthermore, due to VRAM constraints on the NVIDIA A10, we implemented 4-bit QLoRA. While this mixed-precision approach successfully prevented memory exhaustion, the quantization-induced precision loss and gradient noise likely acted as a contributing factor to the modest absolute mAP score. We hypothesize that utilizing higher-bit precision, if hardware had permitted, would yield a noticeable performance gain in coordinate regression. Our data-driven model successfully locates the existence and semantic type of layout elements, struggling primarily with boundary micromanagement compared to traditional discriminative object detectors (Wang et al., 2023).

5. Conclusion

In the EvaHan 2026 competition, we demonstrated that a parameter-efficient, data-centric approach can redefine the state-of-the-art for Ancient Chinese OCR. By strictly using official data alongside a highly optimized Low-Rank Adaptation (Rank 16), we achieved near-human accuracy (97.19% F1) on printed texts, robust parity on handwritten manuscripts, and a 37% mAP enhancement in layout analysis over the official generative baseline. Our work provides strong evidence that coupling explicit document structural priors with Multimodal LLMs is the most effective pathway for end-to-end historical document processing.

While our parameter-efficient framework demonstrates strong overall performance, we acknowledge several limitations that highlight avenues for future research. First, the lack of targeted optimization for handwritten variant characters and degraded glyphs resulted in Task C performance slightly trailing the official baseline. Second, regarding layout analysis, the modest absolute mAP score exposes the inherent constraints of coordinate discretization and quantization-induced precision loss. Furthermore, due to computational resource limitations, comprehensive ablation studies isolating the individual contributions of orientation preprocessing and adaptive prompt engineering were not conducted. Finally, our current approach trains three task-specific adapters independently; future work will explore cross-task generalization to build a unified multimodal model capable of simultaneously handling OCR and layout analysis.

6. Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. We also extend our gratitude to the EvaHan 2026 organizers for providing the datasets and evaluation platform.

7. Bibliographical References

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., ... & Schmidt, L. (2023). OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Chen, T., Saxena, S., Li, L., Fleet, D. J., & Hinton, G. (2021). Pix2seq: A language modeling framework for object detection. *International Conference on Learning Representations*.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Li, B., Feng, M., Hao, X., et al. (2022). Overview of the EvaHan 2022 Shared Task on Ancient Chinese Word Segmentation and POS Tagging. *Proceedings of the LT4HALA 2022 Workshop*.

Li, M., Lv, T., Chen, L., Cui, L., Yin, Y., Hu, W., ... & Wei, F. (2021). TrOCR: Transformer-based optical character recognition with pre-trained models. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, S. Y., et al. (2024). DoRA: Weight-Decomposed Low-Rank Adaptation. *International Conference on Machine Learning (ICML)*.

Liu, Y., et al. (2024). TextMonkey: An OCR-Free Large Multimodal Model for Understanding Document. *arXiv preprint arXiv:2403.04473*.

Masters, D., & Luschi, C. (2018). Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*.

Qiu, L., et al. (2023). Swift: Scalable lightweight infrastructure for fine-tuning. GitHub repository, https://github.com/modelscope/swift.

Wang, J., Jin, L., Ding, K., & Liao, C. (2023). Document pre-training with large language models. arXiv preprint arXiv:2311.03810.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., ... & Lin, J. (2024). Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Yang, Z., et al. (2024). Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*.