A Multi-Modal Recognition Framework for Ancient Books Integrating DoRA-DPO Text Recognition and YOLO Layout Analysis

Chaokun Zhang[1,*], **Xin Wen**[2,*], **Tongtong Zhou**[3,*]
[1]Chinese Classics Research Institute, Fudan University
[2]College of Computer Science and Artificial Intelligence, Fudan University
[3]Center for Historical Geographical Studies, Fudan University
Shanghai, China
{25210100011, 25213050398, 24210760017}@m.fudan.edu.cn

Abstract

The digitization and intelligent analysis of ancient Chinese documents face significant challenges due to diverse scripts, complex layouts, and the prevalence of rare characters. We present a comprehensive multi-modal recognition framework developed for the closed-modality track of the EvaHan 2026 Ancient Chinese Document Multi-Modal Recognition Shared Task. Our approach integrates two specialized pipelines to address these complexities. For text recognition (Tasks A and C), we propose a high-precision OCR system based on the domain-adapted Xunzi_Qwen2_VL_7B_Instruct, leveraging DoRA within a two-stage progressive curriculum learning strategy. To further refine character accuracy, DPO is incorporated alongside a dual-adapter architecture for rare character error localization and correction. For layout detection (Task B), we implement DocLayout-YOLO, enhanced by domain-specific pre-training and Mosaic augmentation to achieve efficient NMS-free element detection. Furthermore, a multi-round robust inference strategy, featuring automatic retry mechanisms and multi-prompt brute-force search, is introduced to handle stubborn and degraded samples effectively. Experimental results demonstrate that our proposed framework achieves superior performance across all evaluation metrics, highlighting its robustness and effectiveness in the digital preservation of ancient Chinese heritage.

Keywords: EvaHan 2026, Ancient Chinese OCR, VLM Fine-tuning, Curriculum Learning, Layout Analysis

1. Introduction

Because ancient Chinese literature constitutes a precious cultural heritage spanning thousands of years, its digital preservation has become a core topic in humanities computing. OCR is key for ancient book digitization, but faces challenges: diverse font styles, rare and variant characters, complex layouts, and degraded image quality.[1]

Recent multimodal large language models (MLLMs) like Qwen2-VL (Wang et al., 2024b) provide new pathways for ancient book OCR. Parameter-efficient methods like LoRA (Hu et al., 2022) and DoRA (Liu et al., 2024) lower adaptation costs. However, MLLMs often struggle with fine-grained spatial coordination and precise element positioning. In the realm of document layout analysis, specialized machine learning frameworks—like those based on the YOLO (Wang et al., 2024a) architecture—remain superior.

We present an integrated multimodal framework specifically optimized for the EvaHan 2026 Track. Our primary contributions are summarized as follows:

- **Curriculum-based DoRA & Dual-Adapter Architecture:** We propose a two-stage progressive fine-tuning strategy to transition knowledge from woodblock prints to handwritten scripts, utilizing DoRA and dual-adapters for efficient cross-domain adaptation.
- **DPO-driven Refinement via Local Tiling:** We apply DPO for character-level correction. Combined with local tiling, this approach significantly enhances precision in recognizing rare and variant characters.
- **Robust Multi-round Inference & Layout Detection:** We implement a high-reliability pipeline featuring automatic retries and multi-prompt search, integrated with a domain-pretrained DocLayout-YOLOv10 for superior end-to-end performance.

2. Related Works

Ancient Book OCR. Traditional methods rely on CNN and sequence modeling (Lombardi and Marinai, 2020). Kim et al. (Kim et al., 2022) introduce Donut, the first OCR-free end-to-end document understanding Transformer. Li et al. (Li et al., 2022)

[*]Authors contributed equally to this work.
[1]The source code for the proposed framework is publicly available at: https://github.com/leeoisaboy/paleomindocr-evahan2026

Proceedings of LT4HALA 2026 @ LREC 2026, pages 268–272.
11 May 2026.

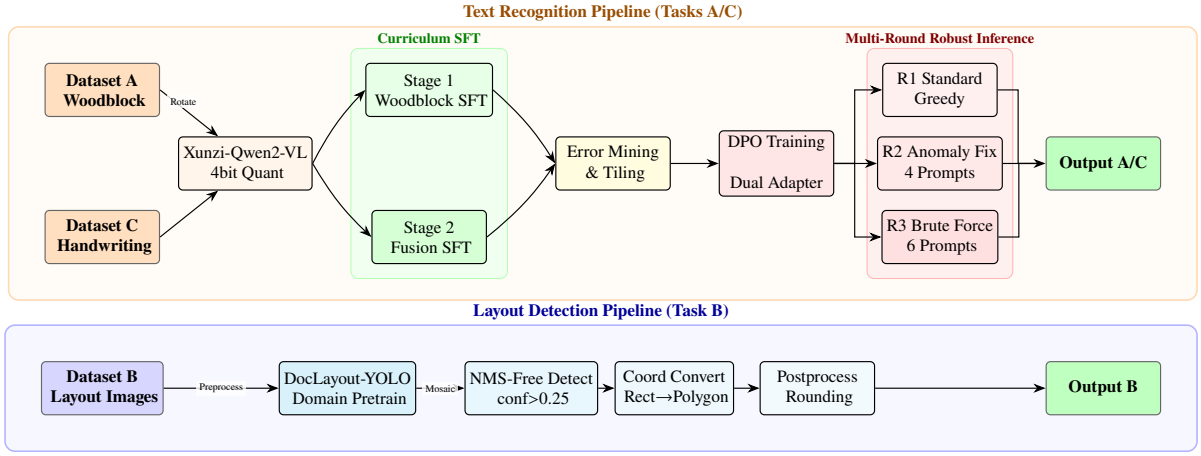

Figure 1: Complete architecture of the closed-modality system

propose DiT, a self-supervised pre-trained document image Transformer. However, VLM-based methods require domain-specific adaptation for ancient books.

Parameter-Efficient Fine-Tuning. Houlsby et al. (Houlsby et al., 2019) introduce Adapter modules. Hu et al. (Hu et al., 2022) propose LoRA, injecting low-rank matrices to reduce memory requirements. Liu et al. (Liu et al., 2024) develop DoRA, decomposing weights into magnitude and direction components for better learning.

Preference Optimization. Ouyang et al. (Ouyang et al., 2022) present RLHF for alignment, but it is complex. Rafailov et al. (Rafailov et al., 2023) introduce DPO, bypassing reward modeling to directly optimize policy models. We apply DPO to character-level OCR error correction.

Document Layout Analysis. Wang et al. (Wang et al., 2024a) propose YOLOv10, eliminating NMS dependency for end-to-end detection. Zhao et al. (Zhao et al., 2024) develop DocLayout-YOLO with domain pre-training. Torres-Aguilar (Aguilar, 2025) shows YOLO variants generalize better on complex layouts.

3. System Approach

3.1. System Overview

The overall architecture of the closed-modality framework is shown in Figure 1, consisting of two independent pipelines.

Text Recognition Pipeline (Tasks A/C): Built upon Xunzi_Qwen2_VL_7B_Instruct, the model undergoes two-stage DoRA curriculum learning SFT, followed by DPO preference optimization for rare character refinement, and finally multi-round robust inference strategy to generate final predictions.

Layout Detection Pipeline (Task B): Based on DocLayout-YOLOv10, initialized with domain-specific pre-trained weights and trained with Mosaic augmentation, achieving fast and accurate layout element localization through confidence-based filtering.

3.2. Text Recognition Pipeline (Tasks A & C)

3.2.1. Base Model and DoRA Configuration

The text recognition pipeline uses Xunzi_Qwen2_VL_7B_Instruct as the base model. The model is based on domain-adapted pre-training on classical Chinese corpora based on Qwen2-VL-7B, providing stronger prior understanding of ancient book characters. We apply 4-bit NF4 quantization with double quantization and BF16 compute precision for training on available hardware.

DoRA adapters cover all Attention layers and MLP layers in the Transformer. Image resolution is uniformly set to 600 px to balance stroke detail preservation and memory constraints. Horizontal strip images from Datasets A and C are rotated 90° clockwise before input to match the traditional vertical reading direction of ancient books.

3.2.2. Two-Stage Progressive Curriculum Learning SFT

We implement a two-stage curriculum strategy to incrementally expand the model’s capabilities.

Stage 1: Woodblock Baseline. Using Dataset A exclusively with specific prompts (e.g., “【雕版印刷】...”), we establish a foundational recognition capability for structured woodblock texts. Training proceeds for 2 epochs with an early stopping mechanism (loss $<$ 0.8 for 10 log entries).

Stage 2: Handwriting Generalization. Dataset C is integrated to introduce complex styles. To balance difficulty, Dataset C is oversampled by $2\times$, re-

sulting in a 1:2 ratio between A and C. This stage employs 2 epochs with a 768-token sequence length to generalize from rigid fonts to fluid handwriting.

Knowledge Retention: This strategy ensures a smooth transition without triggering significant catastrophic forgetting. By employing a mixed-training strategy where woodblock data continuously participates in Stage 2, the model maintains its proficiency in printed styles. Furthermore, DoRA freezes base weights W_0 and only updates the low-rank component ΔW, strictly limiting the overwriting of foundational Stage 1 knowledge. Empirical observations confirm that woodblock recognition accuracy remains stable throughout the handwriting fine-tuning phase.

3.2.3. DPO Preference Optimization for Rare Character Refinement

To rectify persistent errors on rare characters, we introduce DPO through three steps:

(1) Diagnostic Analysis. The Stage 2 model conducts full inference to identify misrecognitions, generating structured error logs pairing ground truth with erroneous predictions.

(2) Tiling and Triples. For each error, we extract a 600×600 pixel local tile. The crop center is fixed at the horizontal center of the rotated image, with the vertical position proportional to the character's text index. Out-of-bounds regions are padded. DPO triples are then constructed: *Prompt* contains the tile; *Chosen* is the ground truth (target character with 3-character context); and *Rejected* is the model's error.

(3) Dual-Adapter Training. Policy and reference adapters are initialized from SFT weights on a 4-bit base. The DPO loss is:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\Big(\beta\big[\log \tfrac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \tfrac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\big]\Big) \quad (1)$$

where y_w is the preferred response, y_l is the rejected response, and $\beta = 0.1$ is the temperature coefficient.

Two key designs for DPO success. (1) an ultra-low learning rate (5×10^{-7}) and tiling to isolate local details from signal dilution; and (2) the frozen reference adapter acting as a semantic anchor, whose implicit KL-divergence constraint ensures localized refinements remain consistent with the global SFT linguistic prior.

3.2.4. Multi-Round Robust Inference Strategy

After DPO training completion, adapter weights are merged into the base model to eliminate PEFT inference overhead, and KV Cache is enabled for acceleration. Inference loads in full precision to maximize recognition accuracy.

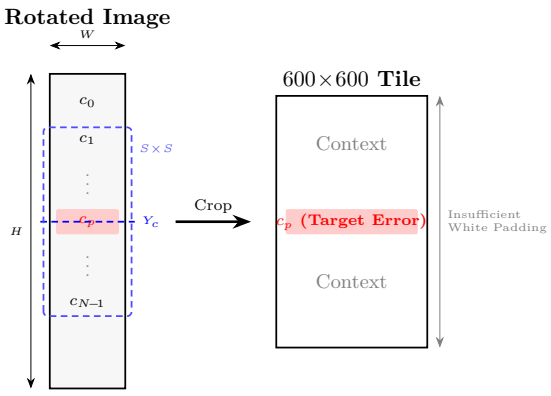

Figure 2: Tiling proportional localization and crop.

To handle various difficult samples, we design a three-round progressive inference strategy:

Round 1 —Standard Inference. Deterministic decoding is performed on all images, retaining only pure CJK characters after post-processing filtering. If output contains invalid keywords or is empty, automatic retry mechanism is triggered to improve success rate.

Round 2 —Anomaly Pattern Specific Prompt Inference. First round results are scanned to identify three anomaly patterns: empty string output, invalid keywords (such as "圖頁" meaning figure page, etc.), and repetitive gibberish (short pattern cycling, long pattern repetition, or extremely few character types in long text). For anomalous samples, 4 different Prompt strategies are used for re-inference, selecting the best non-anomalous output.

Round 3 —Exhaustive Multi-Prompt Inference. For challenging samples, we employ six prompt variants and select the longest valid output to maximize completeness, addressing character omission as the primary failure mode in long-text OCR. To ensure validity, outputs must pass a three-tier diagnostic filter: (1) empty string detection, (2) predefined noise keyword filtering, (3) repetitive gibberish pattern recognition. This heuristic is more robust than majority voting for ancient book corpora, as the high cardinality of the Chinese character set makes identical recognition errors rare, rendering voting-based aggregation less effective. Detailed prompt variants are provided in Appendix A.

The three-round progressive strategy reduces the failure rate of empty output or obvious hallucination in first-round inference to near zero. The six prompt variants describe task requirements from different angles, effectively alleviating model sensitivity to specific wording.

Table 1: Results on Text Recognition Tasks (Tasks A & C).

Task	Models	Consider variant characters				Don't consider variant characters			
		CER	NED	F1	Comp.	CER	NED	F1	Comp.
A	Qwen2.5_VL_7B_Instruct	0.1014	0.0947	0.9110	0.9037	0.1121	0.1054	0.9007	0.8931
	Xunzi_Qwen2_VL_7B_Instruct	0.1786	0.1740	0.8409	0.8282	0.1851	0.1802	0.8345	0.8218
	Ours (DoRA+DPO)	**0.0798**	**0.0795**	**0.9270**	**0.9223**	**0.0872**	**0.0869**	**0.9195**	**0.9149**
C	Qwen2.5_VL_7B_Instruct	0.1207	0.1193	0.8849	0.8812	0.1338	0.1324	0.8718	0.8681
	Xunzi_Qwen2_VL_7B_Instruct	0.1497	0.1492	0.8538	0.8514	0.1609	0.1604	0.8425	0.8402
	Ours (DoRA+DPO)	**0.1322**	**0.1320**	**0.8723**	**0.8692**	**0.1438**	**0.1436**	**0.8609**	**0.8576**

Table 2: Results on Task B (Layout Detection).

Models	mAP@[.5:.95]	Micro F1	Macro F1	IoU
Qwen2.5_VL_7B_Instruct	0	0	0	0
Xunzi_Qwen2_VL_7B_Instruct	0.0236	0.0003	0.0114	0.5943
Ours (YOLO-based)	**0.3889**	**0.6401**	**0.6318**	**0.8119**

4. Experiments

The EvaHan 2026 shared task provides three datasets: Dataset A (woodblock-printed texts), Dataset B (page-level layout annotations), and Dataset C (handwritten texts). Datasets A and C are rotated 90 degrees during training to align with vertical reading direction.

Table 1 presents text recognition results against two baselines without instruction tuning.

Task A: Woodblock Printing. Our system achieves character error rate of 0.0798 (considering variant characters), substantially outperforming Qwen2.5 (0.1014) and Xunzi (0.1786). The comprehensive score reaches 0.9223 with F1 of 0.9270. Without variant consideration, our system maintains 0.0872 error rate and 0.9149 comprehensive score.

Task C: Handwritten Text. Our system achieves 0.1322 character error rate with 0.8692 comprehensive score (considering variants). Qwen2.5 achieves 0.1207 error and 0.8812 score, while Xunzi achieves 0.1497 and 0.8514. Our method shows competitive performance.

Layout Detection. Table 2 shows results. Our YOLO system achieves mAP of 0.3889, micro F1 of 0.6401, and IoU of 0.8119 (37% improvement over Xunzi), substantially outperforming both baselines.

5. Limitations

Despite the robust performance of our framework, several limitations remain. First, the layout detection pipeline (Task B) faces cross-format migration limits; relying on domain-specific pre-training makes it less effective on unconventional layouts like accordion-style bindings or stone rubbings not present in the training set. Second, the text recognition pipeline shows a performance gap on Task C, where highly cursive or severely degraded handwritten scripts challenge the model's understanding of expressive calligraphic strokes. Finally, the multi-round inference strategy improves accuracy at the cost of computational efficiency, introducing latency that limits high-throughput, real-time digitization without further model distillation.

6. Bibliographical References

Sergio Torres Aguilar. 2025. From codicology to code: A comparative study of transformer and yolo-based detectors for layout analysis in historical documents. *arXiv preprint arXiv:2506.20326*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3530–3539.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Francesco Lombardi and Simone Marinai. 2020. Deep learning for historical document analysis and recognition—a survey. *Journal of Imaging*, 6(10):110.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. 2024a. Yolov10: Real-time end-to-end object detection. *Advances in neural information processing systems*, 37:107984–108011.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception.

A. Core Recognition Prompts

Our system employs a hierarchical set of instructions to ensure robust recognition. To avoid character encoding issues and maintain visual consistency, all prompts are delivered in clear, instruction-oriented Chinese.

- **Standard Prompt:** "请仔细观察这张古籍图片，从上到下、从右到左，一个字一个字地识别并输出。要求：1. 每识别一个字就输出一个字；2. 只输出汉字，不要输出任何标点或说明；3. 如果某个字看不清，根据上下文推断；4. 绝对不要输出' 圖頁' 这两个字。"
- **Retry Prompt:** "这张图片包含古籍文字内容（不是空白页或图页标记）。请仔细辨认图中的每一个汉字，按阅读顺序逐一输出。输出格式：直接输出连续的汉字，无标点无空格。"
- **Anomaly Repair Prompts:**
  1. "请逐字识别图中的古籍文字，只输出文字，不要输出任何其他内容。"
  2. "这是一张古籍图片，请仔细辨认每一个汉字并输出，只输出识别到的文字。"
  3. "请识别图片中的所有汉字，按照从右到左、从上到下的顺序输出纯文字。"
  4. "图中是古籍文字，请精确识别并输出，注意笔画细节，只输出汉字。"