Multimodal Ancient Document Parsing: Technical Report for EvaHan2026 Competition

Liqi He[1], **Qiwei Li**[2], **Ziye Yang**[3], **Zuchao Li**[2]
[1]School of Computer Science, Wuhan University, Wuhan, 430072, China,
[2]School of Artificial Intelligence, Wuhan University, Wuhan, 430072, China,
[3] College of Information and Intelligence, Hunan Agricultural University
heliqi@whu.edu.cn, qw-line@whu.edu.cn, yzy_0607@outlook.com, zcli-charlie@whu.edu.cn

Abstract

We present the multimodal Optical Character Recognition (OCR) and layout analysis methods developed for the EvaHan 2026 competition. Our approach is built upon the Qwen2.5-VL-7B-Instruct architecture and integrates two core strategies: (1) a reinforcement learning alignment pipeline utilizing Direct Preference Optimization (DPO) and Group Relative Policy Optimization (GRPO) to explicitly mitigate hallucination and coordinate instability; and (2) a four-stage curriculum learning framework that synthesizes domain-specific historical artifacts to enhance open-modality generalization. Using this approach, we achieve competitive results, notably reaching a Character Error Rate (CER) of 0.0303 on printed texts (Task A) and 0.0552 on handwritten manuscripts (Task C), as well as an Average Intersection over Union (IoU) of 0.7638 on layout element analysis (Task B).

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

1. Introduction

The digitization of ancient Chinese literature is a critical endeavor for the preservation and dissemination of cultural heritage. Vast volumes of handwritten and archaic printed historical documents remain fragile and inaccessible to modern computational analysis. Optical Character Recognition (OCR) technology serves as the foundational bridge to transform these imaged texts into structured, editable data. To advance this field, the Fifth International Evaluation of Ancient Chinese Information Processing (EvaHan 2026) has introduced challenging multimodal benchmarks focusing on printed text recognition (Task A), layout element analysis (Task B), and handwritten character recognition (Task C).

Despite the recent success of Vision-Language Models (VLMs) in general-domain document AI, applying these models directly to ancient Chinese texts presents significant challenges. Printed historical texts (Task A) frequently contain rare variant characters, complex woodblock layouts, and physical degradation. Handwritten manuscripts (Task C) introduce even greater variability through cursive connections, stroke omissions, and highly personalized calligraphic styles. Furthermore, for layout analysis (Task B), VLMs often struggle with fine-grained spatial perception, leading to coordinate hallucinations when detecting overlapping elements such as seals superimposed on text blocks.

To address these challenges, we present a robust, unified multimodal framework built upon the well-established Qwen2.5-VL-7B-Instruct architecture. Since supervised fine-tuning (SFT) struggles with VLM hallucinations in dense transcription and coordinate prediction, we propose an reinforcement learning (RL)-aligned and data-scaled training pipeline.

For the **closed modality**, where external data is strictly prohibited, we introduce a progressive preference alignment pipeline. We leverage Direct Preference Optimization (DPO) to penalize the prediction of visually similar but incorrect variant characters (hard negatives). Most notably, we introduce the application of Group Relative Policy Optimization (GRPO) in ancient OCR. By designing customized reward functions targeting character-length consistency and geometric Intersection-over-Union (IoU), our GRPO strategy effectively mitigates the coordinate hallucination problem in layout analysis and reduces character omission rates.

For the **open modality**, we propose a **four-stage curriculum learning** pipeline to systematically build the model's visual-semantic representations. By progressing from single-character recognition using massive multi-font libraries to domain-specific synthesized context learning (e.g., simulating woodblock prints and cursive stroke erosion), we provide a smooth learning trajectory that maximizes the model's capacity to filter out historical noise before integrating real-world data.

Our experiments demonstrate the superiority of our proposed framework. In the official EvaHan 2026 evaluation, our system achieved competitive results across all tasks, notably reaching a Character Error Rate (CER) of 0.0303 in Task A and 0.0552 in Task C, while our RL-driven spatial standardization elevated the layout analysis performance in Task B.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 322–329.

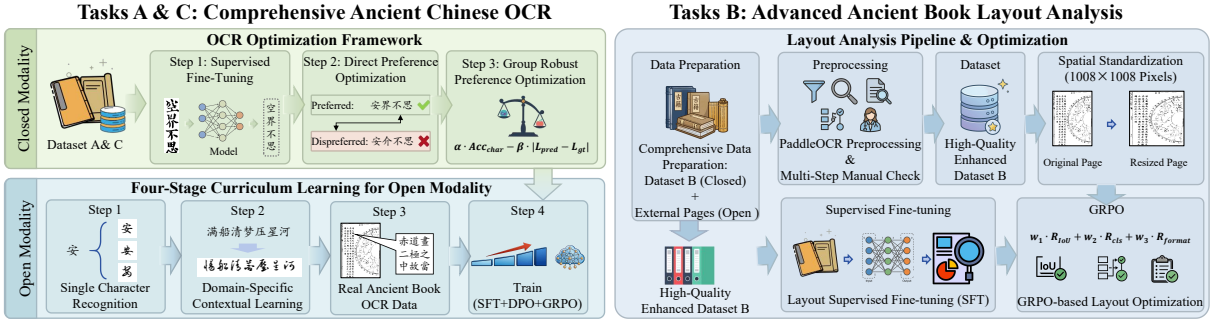

Figure 1: Overview of our methodology.

2. Related Work

2.1. Ancient Chinese Text Processing and Digitization

The digitization of ancient Chinese texts faces unique challenges due to complex layout structures and degraded characters (Li et al., 2024a). Recent advancements have shifted toward domain-specific foundation models, such as GujiBERT (Wang et al., 2023), SongSong (Hu et al., 2025) and the Xunzi large language model series (Zhu et al., 2025), to capture the contextual semantics of classical Chinese. However, end-to-end multimodal digitization for handwritten and woodblock texts remains challenging due to stroke overlapping and ink diffusion.

2.2. Document AI and Vision-Language Models

Early Document AI systems typically relied on complex cascaded pipelines, requiring separate, specialized models for text detection (Du et al., 2020), layout parsing (Huang et al., 2022), and key information extraction (Li et al., 2024b), which often led to error propagation across stages. In contrast, modern VLMs like Donut (Kim et al., 2022) and Nougat (Blecher et al., 2023) have significantly advanced the field by treating document understanding as an end-to-end image-to-text generation task, eliminating the need for separate OCR components. Furthermore, recent large-scale models such as LLaVA (Liu et al., 2023) and the Qwen-VL series (Bai et al., 2023, 2025) have achieved substantial improvements in performance by integrating high-resolution visual encoders with powerful language decoders, enabling the fine-grained perception of both textual content and spatial layout structures.

2.3. Preference Alignment and Reinforcement Learning

Standard SFT adapts VLMs well but suffers from exposure bias and hallucinations, like character omission or coordinate fabrication. To align model outputs with human preferences, reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) has become a standard protocol. DPO (Rafailov et al., 2023) simplifies this by optimizing the policy model directly on preference pairs. Furthermore, GRPO (Shao et al., 2024) and curriculum learning pipelines (Bengio et al., 2009) have emerged to efficiently optimize discrete sequence generation and mitigate hallucinations in multimodal layout analysis.

3. Methodology

In this section, we detail our technical approach for the EvaHan 2026 competition, focusing on ancient Chinese OCR (Tasks A & C) and layout element analysis (Task B). Our solution leverages the well-established Qwen2.5-VL-7B-Instruct as the foundational VLM for all tasks. This 7B-parameter scale was selected to adhere to the competition's uniform requirements, ensuring a standardized and fair baseline for evaluating inference latency and practical deployment overhead among all participants.

3.1. Tasks A & C

Tasks A and C focus on printed and handwritten ancient Chinese text recognition, respectively. The primary challenges involve identifying rare/variant characters and handling stroke blurring in manuscripts.

3.1.1. Closed Modality

Restricted to official training data, our closed modality strategy refines precision via post-training alignment:

SFT: The model is fine-tuned on the official Dataset A and C.

DPO: We extract hard negative samples where the baseline model misidentified characters. Using

DPO, we train the model to distinguish between the correct transcription (chosen) and the common errors (rejected).

GRPO: To address the common VLM issue of character omission or hallucination, we employ GRPO. We design a reward function that balances character accuracy and sequence length:

$$R = \alpha \cdot Acc_{char} - \beta \cdot |L_{pred} - L_{gt}| \quad (1)$$

where Acc_{char} denotes the accuracy of character recognition, L_{pred} and L_{gt} denote the predicted and ground truth lengths, respectively.

3.1.2. Open Modality

The open modality allows the use of external resources. We implement a **four-stage curriculum learning** pipeline:

Stage 1 (Single Character Recognition): We utilize a large-scale library of traditional Chinese characters, rendered with diverse ancient-style fonts (e.g., Cao, Xing, and Kai styles) to build robust visual feature extraction for individual glyphs.

Stage 2 (Domain-Specific Contextual Learning): We synthesize long-context data using ancient Chinese corpora. For Task A, we use Kai-style fonts to simulate woodblock printing; for Task C, we use cursive (Cao/Xing) fonts with stroke erosion and overlapping to simulate manual handwriting.

Stage 3 (Real OCR Data): We integrate authentic OCR datasets sourced from the internet and open-source historical archives to bridge the gap between synthetic distributions and real-world document characteristics.

Stage 4 (Supervised Fine-tuning & Reinforcement learning): The model is fine-tuned on the official Dataset A and C. We apply the same DPO and GRPO strategies used in the closed modality to achieve final convergence.

3.2. Task B

Task B focuses on the identification of four critical layout elements: *text*, *image*, *book_edge*, and *seal*. To address the challenges of domain adaptation and overlapping categories, we implement a unified pipeline encompassing annotation enhancement, spatial standardization, and structural reinforcement.

3.2.1. Data Preparation

The core distinction between our closed and open modality solutions lies in the data acquisition and annotation scale.

Closed Modality: We focus exclusively on the official Dataset B. To improve label quality, we perform *annotation enhancement* by utilizing PaddleOCR for preliminary detection followed by a manual audit. This process emphasizes the correction of fine-grained boundaries and the resolution of label conflicts in overlapping regions, such as seals superimposed on text blocks.

Open Modality: In addition to the enhanced official dataset, we introduce supplementary ancient book pages gathered from external digital archives. These additional samples undergo the same "PaddleOCR + manual review" pipeline to ensure high-quality regions annotations.

3.2.2. Spatial Standardization

To establish a stable coordinate mapping system across heterogeneous document sizes, all input images are resized to a fixed resolution of 1008x1008 pixels. This standardized square resolution not only ensures optimal alignment with the vision-language backbone's patch processing mechanism, but also provides pixel-level granularity for detecting small elements like seals and thin book edges, effectively balancing fine-grained spatial perception with computational efficiency.

3.2.3. SFT

Using the processed annotations, we conduct supervised fine-tuning on the Qwen2.5-VL-7B-Instruct model. The training objective is to map visual features to structured layout descriptions, enabling the model to master basic category classification and bounding box localization within a normalized relative coordinate system, where all spatial positions are scaled to a standardized range of $[0, 1000]$.

3.2.4. Reinforcement Learning with Structural Rewards

To further refine the geometric precision and format stability, we employ GRPO. We define a multi-dimensional reward function to supervise the model's output:

Geometric IoU Reward: Rewards are calculated based on the IoU between predicted bounding boxes and ground truth, directly optimizing pixel-level localization accuracy.

Table 1: Official Results for Task A (Printed) and Task C (Handwritten). Variants: "Yes" indicates considering variant characters, "No" indicates ignoring them.

Task	Method	Consider Variant (Yes)				Consider Variant (No)			
		CER ↓	NED ↑	F1 ↑	Score ↑	CER ↓	NED ↑	F1 ↑	Score ↑
Task A	Base Model	0.1014	0.0947	0.9110	0.9037	0.1121	0.1054	0.9007	0.8931
	Instruction Tuning	0.0618	0.0613	0.9430	0.9397	0.0685	0.0679	0.9364	0.9331
	Closed (Ours)	0.0316	0.0314	0.9702	0.9690	0.0316	0.0314	0.9702	0.9690
	Open (Ours)	**0.0303**	**0.0301**	**0.9715**	**0.9703**	**0.0303**	**0.0301**	**0.9715**	**0.9703**
Task C	Base Model	0.1207	0.1193	0.8849	0.8812	0.1338	0.1324	0.8718	0.8681
	Instruction Tuning	0.0920	0.0919	0.9099	0.9086	0.1066	0.1065	0.8953	0.8940
	Closed (Ours)	0.0751	0.0750	0.9267	0.9255	0.0902	0.0901	0.9116	0.9104
	Open (Ours)	**0.0552**	**0.0549**	**0.9471**	**0.9456**	**0.0703**	**0.0700**	**0.9320**	**0.9305**

Classification Reward: Ensures the predicted labels correctly identify the element type (e.g., distinguishing between a text block and an image).

Format Consistency Reward: To mitigate the common "hallucination" of coordinates in VLMs, we impose penalties on outputs that violate the structured format (e.g., `<LABEL>[x1,y1,x2,y2]`), ensuring the results are programmatically robust.

4. Experiments

4.1. Experimental Settings

Model training was distributed across two NVIDIA A40 GPUs (48GB VRAM) using LoRA fine-tuning ($r = 64$, $\alpha = 128$). We set the per-device training batch size to 4 with 16 gradient accumulation steps. For reproducibility, models for Tasks A and C were trained for 1 epoch, while the Task B model was trained for 2 epochs. All inference was conducted on a single A40 GPU with the temperature set to 0 (for determinism) and a maximum context length of 4096 tokens to support high-resolution image patches and long sequence transcriptions.

4.2. Main Results on EvaHan 2026 Official Test Sets

The main results reflect our performance on the official EvaHan 2026 evaluation system. We report results for both the **closed modality** (strictly adhering to official models and data) and the **open modality** (leveraging our four-stage curriculum learning and external synthetic/real-world data).

OCR Tasks (Tasks A & C): Table 1 summarizes the performance on Tasks A and C. In Task A (printed texts), our system achieved a comprehensive score of 0.9703, outperforming the instruction-tuned baseline. In Task C (handwritten texts), the advantage of our **open modality** is even more pronounced, reaching a CER of 0.0552, which is a 40% relative improvement over the instruction-tuned baseline (0.0920).

Table 2: Official Evaluation of Task B.

Method	mAP@.5	Micro F1	Macro F1	Avg IoU
Base Model	0.0000	0.0000	0.0000	0.0000
Instruction Tuning	0.2006	0.0513	0.1530	0.6600
Closed (Ours)	**0.5941**	**0.8342**	**0.8686**	**0.7638**
Open (Ours)	0.5818	0.8278	0.8645	0.7572

Layout Element Analysis (Task B): As shown in Table 2, the base VLM fails on coordinate-based layout tasks (mAP@.5 of 0). However, our SFT and subsequent GRPO refinement for spatial standardization enable the model to reach a Micro F1 of 0.8342. Based on our analysis, the slight performance drop of the open modality on Task B can be attributed to annotation discrepancies. Upon inspecting the official training data, we identified inconsistencies and errors in the layout labels. To mitigate this, we performed additional manual re-annotation and introduced self-labeled data in the open modality setting. A detailed visual analysis of this granularity mismatch is provided in Appendix B. While this data provides richer spatial diversity, its annotation logic partially diverges from the official evaluation standards, leading to a marginal performance gap compared to the closed modality, which strictly adheres to the original official annotations. Beyond manual re-annotation, future mitigation strategies will involve applying rule-based post-processing filters to merge fine-grained detections into official macro-level blocks, or adopting a multi-granularity evaluation metric.

4.3. Ablation Study

To evaluate the contributions of our proposed alignment strategies (DPO, GRPO) and curriculum data generation, we conducted an ablation study on a local validation split. Because the official test sets were strictly withheld, we partitioned the official training datasets (A, B, and C). For each task, we randomly sampled 500 image-text pairs to form a

Table 3: Ablation results on Task A and Task C.

Method	Field	Task A		Task C	
		CER ↓	F-Score ↑	CER ↓	F-Score ↑
Base Model	-	0.1420	0.8741	0.3850	0.7253
SFT (Direct Tuning)	Closed	0.0180	0.9837	0.1119	0.8912
+ DPO	Closed	0.0176	0.9840	0.0907	0.9122
+ **GRPO**	Closed	0.0169	0.9848	0.0898	0.9128
+ Random Synthetic Data	Open	0.0177	0.9841	0.1063	0.8963
+ Ancient Books Syn.	Open	0.0172	0.9850	0.0924	0.9106
+ **Real-world Full**	Open	**0.0161**	**0.9855**	**0.0770**	**0.9246**

Table 4: Ablation results on Task B.

Method	Field	mAP@.5 ↑	Micro F1 ↑	Macro F1 ↑
SFT (Direct Tuning)	Closed	0.5234	0.7140	0.6968
+ GRPO	Closed	0.7668	0.8935	0.9357
+ External Data	Open	**0.8028**	**0.9788**	**0.9842**

held-out local test set, preserving the original distribution of document layouts, font variants, and degradation artifacts.

All baseline and intermediate models were trained on the remaining data and evaluated exclusively on this 500-sample set. This localized setup ensures a fair comparison to accurately isolate the performance gains achieved by transitioning from standard SFT to advanced alignment techniques in the closed modality, as well as the incremental benefits of integrating domain-specific external resources in the open modality.

Efficacy of RL Strategies (Cumulative Alignment): The transition from the base model to our final closed modality submission demonstrates a clear cumulative benefit from our alignment pipeline. As shown in Table 3, standard SFT provides the initial domain adaptation, reducing the CER (e.g., from 0.3850 to 0.1119 in Task C). Building upon this SFT foundation, the introduction of DPO teaches the model to reject visually similar hard negative characters, further decreasing the CER to 0.0907. Finally, integrating GRPO on top of the SFT+DPO checkpoint yields the best closed-domain performance (CER 0.0898). The impact of GRPO is more striking in Task B (Table 4), where applying our multi-dimensional structural rewards on top of the SFT model causes the mAP@.5 to increase from 0.5234 to 0.7668. This improvement demonstrates that geometric and format-consistency rewards can effectively mitigate the inherent coordinate hallucination issues in standard VLMs.

Impact of Domain-Specific Data (Open Modality Escalation): Our open modality experiments incrementally introduced external data on top of the established RL framework, revealing critical insights into data scaling for ancient texts. Interestingly, when we naively added **random synthetic data**, performance actually degraded compared to our best closed-domain GRPO model (e.g., in Task C, CER worsened from 0.0898 to 0.1063). This indicates that when dealing with diverse and variable handwritten styles such as calligraphy, the randomness introduced by generic synthetic characters disrupts the contextual comprehension that the model heavily relies on for accurate judgment. However, when we transitioned to our stage 2 **ancient books synthetic data**—which accurately simulates woodblock printing and cursive stroke erosion within coherent textual contexts—the model successfully integrated the external knowledge, recovering and improving its robustness (task C CER dropping to 0.0924). Ultimately, combining the full curriculum with high-quality **real-world data** advanced the system to its peak performance across all tasks (task A CER 0.0161, task C CER 0.0770). This trajectory confirms that while synthetic curriculum learning bridges the initial glyph recognition gap, real-world domain diversity remains crucial for handling complex historical noise.

5. Future Work

In future work, we plan to explore end-to-end multi-task architectures capable of jointly performing layout element analysis and character transcription in a single forward pass. Additionally, we aim to leverage advanced generative models (e.g., diffusion models) to synthesize more realistic historical degradation artifacts (such as ink bleeding and paper aging), further minimizing the domain gap in our curriculum learning pipeline.

6. Conclusion

We presented a comprehensive solution for EvaHan 2026 using Qwen2.5-VL-7B-Instruct. Our RL alignment pipeline (DPO and GRPO) effectively mitigates VLM hallucinations and coordinate instability. Coupled with a four-stage curriculum learning strategy, our system demonstrated robust performance across printed texts, handwritten manuscripts, and complex layouts, achieving competitive results in both the closed and open modalities.

7. Limitations

Despite strong performance, our RL phases (especially GRPO) are computationally expensive and hyperparameter-sensitive (Appendix A). Additionally, Task B's open modality annotation enhancement relies on manual auditing, restricting scalability on massive unlabeled corpora. Appendix C details remaining challenges with rare variant characters, including visual confusion and modern substitution.

8. 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 understanding, localization. *Text Reading, and Beyond*, 2(1):1.

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*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

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

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*.

Jiliang Hu, Jiajia Li, Ziyi Pan, Chong Chen, Zuchao Li, Ping Wang, and Lefei Zhang. 2025. Songsong: A time phonograph for chinese songci music from thousand of years away. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26229–26237.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.

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.

Bin Li, Bolin Chang, Zhixing Xu, Minxuan Feng, Chao Xu, Weiguang Qu, Si Shen, and Dongbo Wang. 2024a. Overview of evahan2024: The first international evaluation on ancient chinese sentence segmentation and punctuation. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 229–236.

Qiwei Li, Zuchao Li, Ping Wang, Haojun Ai, and Hai Zhao. 2024b. Hypergraph based understanding for document semantic entity recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2950–2960.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, et al. 2023. Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts. *arXiv preprint arXiv:2307.05354*.

Dongmei Zhu, Chang Liu, Xue Zhao, Zhixiao Zhao, Si Shen, and Dongbo Wang. 2025. Xunzi-mllm: a multimodal large language model for ancient text and image recognition. *Digital Scholarship in the Humanities*, 40(2):709–722.

A. Computational Cost Analysis

To address the computational overhead introduced by the reinforcement learning alignment pipeline, we explicitly quantify the training time and GPU memory footprint. Table 5 details the resource consumption for Task A (closed modality) trained on NVIDIA A40 GPUs (48GB VRAM).

As shown in the table, while supervised fine-tuning (SFT) is highly efficient, requiring approximately 1.8 hours, the alignment phases significantly increase the computational burden. Direct Preference Optimization (DPO) extends the duration to roughly 12.8 hours. Most notably, Group Relative Policy Optimization (GRPO) demands over 32 hours and consumes approximately 45GB of VRAM, approaching the hardware limit. This substantial $18\times$ increase in training time compared to SFT is primarily attributed to the group sampling mechanism, which necessitates generating and evaluating multiple candidate trajectories per prompt to compute relative rewards. To mitigate scalability concerns, our GRPO pipeline is exclusively applied to a small, targeted subset of hard-negative examples rather than the entire massive dataset, ensuring computational feasibility.

Furthermore, the selection of GRPO over standard Proximal Policy Optimization (PPO) (Schulman et al., 2017) is driven by both algorithmic suitability and hardware constraints. From an algorithmic perspective, PPO relies on a value model to predict absolute expected returns, which typically struggles to accurately map highly discrete, rule-based, and structural reward functions (such as rigid coordinate format matching and geometric IoU calculation in Task B). Conversely, GRPO naturally accommodates these non-differentiable metrics by applying relative reward normalization within a sampled group of outputs. This mechanism directly evaluates the relative advantage of candidate coordinate sequences and transcriptions without requiring absolute value estimations, thereby mitigating the optimization instability inherent in Actor-Critic architectures.

Practically, this elimination of the separate value model—which typically mirrors the 7B-parameter scale of the policy model—is also a critical enabler. It confines the peak memory footprint to approximately 45GB per GPU, rendering the reinforcement learning phase computationally feasible on our dual NVIDIA A40 hardware, whereas standard PPO would likely trigger out-of-memory (OOM) errors.

Table 5: Computational Cost for Task A (Closed Modality) on NVIDIA A40 (48GB).

Training Phase	Peak VRAM	Training Time (Hours)
SFT	∼28 GB	1.81
DPO	∼40 GB	12.84
GRPO	∼45 GB	32.56

B. Analysis of Annotation Discrepancies in Task B

To address the slight performance degradation observed in Task B (layout element analysis), we provide a visual comparison that highlights the fundamental discrepancies in annotation logic between the official evaluation standards and our re-optimized annotation pipeline.

As illustrated in Figure 2, the official dataset often employs a macro-level annotation strategy for complex images such as historical maps. Figure 2(a) shows that only the major vertical text columns are annotated, while scattered, fine-grained text elements (e.g., small geographical labels like town or river names) are entirely ignored.

Conversely, to train a more robust and comprehensive layout analysis model, our optimized annotation pipeline utilized PaddleOCR combined with manual auditing to perform exhaustive, micro-level annotation. As shown in Figure 2(b), our enhanced annotations explicitly capture all visible text elements regardless of their size or location.

Consequently, when the model trained with our re-optimized annotations is evaluated against the official ground truth, its ability to successfully detect these fine-grained elements works to its detriment. The official evaluation metric penalizes these correct (but unannotated in the ground truth) predictions as false positives. This systematic mismatch in annotation granularity—rather than a deficiency in model capability—is the primary cause of the slight drop in the mAP@.5 score (from 0.5941 to 0.5818).

C. Failure Case Analysis for Rare and Variant Characters

To address the remaining challenges, we conducted a failure case analysis focusing on extremely rare and variant characters. While our pipeline significantly reduces the CER, standard VLMs still exhibit specific failure patterns, categorized into two primary types as illustrated in Figure 3:

1. Visual Feature Confusion (Figure 3, Left): The model incorrectly predicts a visually similar common character instead of the rare ground truth character. In historical woodblock prints, degraded ink and compact stroke structures often obscure

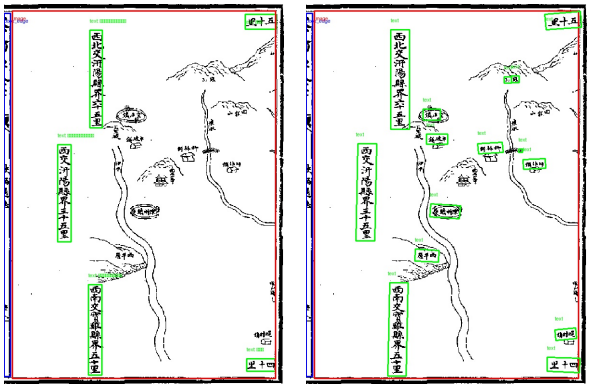

Figure 2: Comparison of annotation logic for Task B. The official dataset (left) ignores scattered fine-grained text on maps, whereas our re-optimized annotations (right) exhaustively label all text elements.

fine-grained radical differences, leading the VLM's visual encoder to make shape-based misclassifications.

2. Variant Character Substitution (Figure 3, Right): The model predicts the standard modern character instead of its historical variant counterpart. Because the foundation model's language decoder is heavily pre-trained on modern standard text, its strong linguistic prior can occasionally override the visual evidence, causing an "auto-correction" effect that bypasses the strict literal transcription required for this task.

Figure 3: Typical failure cases. Left: Visual confusion due to structural similarity. Right: Variant character substitution driven by the VLM's modern linguistic priors.