Data-Centric Strategies for Ancient Chinese Text Recognition: Augmentation, Annotation Refinement, and Style Transfer in EvaHan 2026

Chengfei Li[1] , Yunjie Zhang[1] , Xiaoyi Li[1] , Changshun Quan[1] , Taihe Cao[1] , Bin Liu[1]

[1] Artificial Intelligence Research Institute on Education, Qilu Normal University, China
lichengfei@qlnu.edu.cn; 18769301321@163.com; 15698042890@163.com; 19707032403@163.com; cth@qlnu.edu.cn; liubin@qlnu.edu.cn

Abstract

This paper describes our system for the EvaHan 2026 shared task. We design and experiment with data-centric strategies across three subtasks: printed text OCR (Task A), layout element analysis (Task B), and handwritten text OCR (Task C). Our approach employs systematic data augmentation using 17 transformation strategies, comprehensive manual annotation refinement for layout analysis, and style transfer augmentation for handwritten texts. We use pre-trained Qwen2.5-VL-7B-Instruct with LoRA fine-tuning as the base model. According to the evaluation metrics adopted by the organizers, our system achieves 27.5% and 4.5% CER reduction over official baselines for Tasks A and C respectively. Manual annotation refinement for Task B achieves 205% improvement in Micro F1 and 258% improvement in Macro F1 on the validation set, demonstrating that annotation quality is the primary bottleneck for layout analysis in closed-modality settings.

Keywords: ancient Chinese OCR, data augmentation, annotation refinement, multimodal language models

1. Introduction

Ancient Chinese texts represent invaluable cultural heritage, documenting thousands of years of civilization through printed books, hand-copied manuscripts, and administrative documents. However, their automated processing remains challenging due to several factors: character variants across different dynasties and regions, material degradation from aging and environmental damage, complex multi-column layouts with embedded illustrations and seals, diverse calligraphic styles reflecting individual scribal traditions, and critically, the scarcity of high-quality annotated training data. The EvaHan 2026 evaluation campaign aims to assess the capabilities of large language models in addressing these challenges across three dimensions: printed text recognition (Task A), layout element analysis (Task B), and handwritten text recognition (Task C).

We design and experiment with three data-centric approaches. The systematic data augmentation approach expands training data through photometric and geometric transfor-mations simulating realistic degradation patterns. The annotation refinement approach manually corrects layout element labels to address systematic errors including missing text regions. The style transfer approach generates artistic variants for handwritten texts using traditional Chinese artistic styles.

We participate exclusively in the closed track, which restricts participants to official training data and two specified models: Qwen2.5-VL-7B-Instruct and Xunzi_Qwen2_VL_7B_Instruct. This constraint forces data-centric optimization rather than model scaling.

According to the evaluation metrics adopted by the organizers, our system achieves competitive performance across all three tasks. Compared to the baseline Qwen2.5-VL-7B-Instruct, our augmentation strategies obtain substantial improvements for Tasks A and C. Our annotation refinement achieves dramatic gains for Task B on validation data, though test performance reveals annotation quality issues in the original dataset. Our code and annotation tool are available at https://github.com/SeekingYourRoots/EvaHan2026-TaskB-Annotation-Tool.

2. Related Work

Ancient Document OCR Deep learning has revolutionized historical OCR through end-to-end architectures. Shi et al. (2017) proposed CRNN combining CNN-RNN-CTC for sequence recognition. Attention-based models improved irregular layout handling (Cheng et al., 2017). Recent vision-language models like TrOCR (Li et al., 2023) and Donut (Kim et al., 2022) leverage large-scale pre-training for document understanding.

Data Augmentation Simard et al. (2003) demonstrated elastic distortions improve handwriting recognition. Buslaev et al. (2020) introduced Albumentations for fast augmentation pipelines. Cubuk et al. (2019) proposed AutoAugment using reinforcement learning to discover optimal strategies.

Layout Analysis Modern approaches employ deep object detection. Faster R-CNN (Ren et al., 2015) and Mask R-CNN (He et al., 2017) enable accurate element localization. MMDetection (Chen et al., 2019) provides unified implementtations for document layout analysis.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 305–310.

3. Our Method

In this section, we introduce our methods and model architectures for the three subtasks.

The EvaHan 2026 task encompasses three subtasks operating on different data modalities. Task A (printed texts) benefits from systematic photometric degradation simulation. Task B (layout analysis) requires high-quality spatial annotations. Task C (handwritten texts) exhibits calligraphic style constraints limiting augmentation intensity.

We employ Qwen2.5-VL-7B-Instruct as the base model across all tasks, using LoRA for parameter-efficient fine-tuning. This unified architecture isolates the effects of data interventions.

3.1 Data Augmentation Pipeline

We implement 17 transformation strategies using Albumentations organized into four categories:

Color/Lighting : RandomBrightnessContrast ($\pm$ 20%), RandomGamma ($\gamma \in$ [0.7,1.3]), HueSaturationValue (H $\pm$ 10°, S $\pm$ 30, V $\pm$ 20), RGBShift ($\pm$20), and combined color jittering.

Blur: GaussianBlur (kernel 3-7, $\sigma \in$ [0.1,2.0]), MotionBlur (kernel 3-7), MedianBlur (kernel 5), and average blur.

Noise: GaussNoise ($\sigma \in$ [10,50]), JPEG compression (quality 50-80), PixelDropout (1%), and MultiplicativeNoise ($\times$0.9-1.1).

Mixed: Combined transformations including brightness+noise, gamma+blur, motion+compression, and HSV+noise+blur.

All transformations preserve original image dimensions, which is critical for maintaining bounding box annotation validity in Task B. The pipeline randomly selects one transformation per augmented image.

3.2 Annotation Refinement for Task B

Analysis of validation results revealed three systematic annotation errors:

Missing text regions: Substantial textual content present but not annotated, causing false positive penalties when models correctly detect unlabeled regions.

Element type confusion: Ambiguous bound-aries between book_edge/text, image/seal, and seal/text categories.

Boundary inaccuracies: Bounding boxes insufficiently covering element extents, reducing IoU scores.

We developed a web-based annotation tool (Figure 1) to facilitate manual correction. The tool features intuitive dual drawing modes, color-coded element types for clear distinction, efficient keyboard shortcuts, and real-time persistence of annotation data.

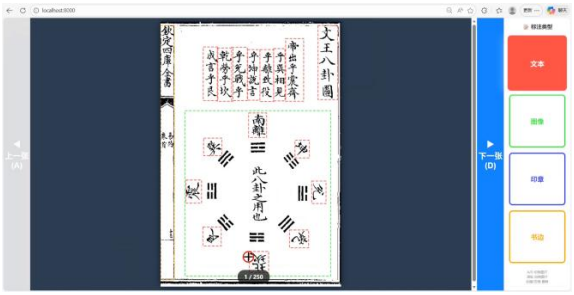

Figure 1: Web-based annotation tool interface showing layout element correction capabilities

We manually corrected all 5,000 Dataset B training images. The correction process focused primarily on adding missing text region annotations (approximately 68% of corrections) and resolving element type confusion (approximately 24% of corrections).

3.3 Style Transfer for Task C

We distinguish three traditional Chinese artistic styles and generate variants (Figure 2) using the Qwen-Image-Edit API:

Stone Inscription: Bold angular strokes with weathering effects simulate ancient stone tablet carvings. Prompt: "Transform this handwritten text into stone inscription style, as if carved on an ancient Chinese stone tablet, maintaining all character content exactly."

Wood Carving: Raised relief with wood grain texture emulates woodblock printing. Prompt: "Convert this handwritten text to wood carving style, as if carved on traditional Chinese wooden blocks, keeping all text content unchanged."

Ink Wash: Varying stroke thickness with ink diffusion effects is characteristic of traditional calligraphy. Prompt: "Rerender this handwritten text in traditional ink wash painting style using classical Chinese brush techniques while preserving all character forms."

Manual quality filtering is implemented to ensure content preservation, with approximately 8-12% of generated images requiring removal due to character modification errors or illegibility introduced during style transfer.

Figure 2: Style transfer examples showing the three traditional Chinese artistic styles: Stone Inscription with bold angular strokes (left), Wood Carving displaying raised relief texture (middle), and Ink Wash exhibiting stroke thickness variation and ink diffusion (right).

3.4 Model Architecture

Our model architecture employs Qwen2.5-VL-7B-Instruct as the vision-language backbone. The model processes image-text pairs through:

Vision Encoder: Processes input images to extract visual features.

Language Model: Qwen2.5-7B transformer with 28 layers, 3584 hidden dimensions.

LoRA Adaptation: Low-rank adaptation matri-ces targeting query and value projection layers for parameter-efficient fine-tuning.

For Task B, the model outputs bounding box coordinates and element type classifications. For Tasks A and C, the model generates text sequences representing recognized characters.

4. Experiments

4.1 Data

The official training data originates from digitized ancient Chinese texts:

Dataset A (Printed Texts) consists of data selected from the Siku Quanshu (Complete Library of the Four Treasuries), including classics, history, philosophy, and literature, as well as various other ancient books.

Dataset B (Mixed Layouts) contains mixed image-text data selected from the Siku Quanshu and other ancient books.

Dataset C (Handwritten Texts) includes hand-written ancient books, primarily the Chinese Buddhist canon, including the Chinese Buddhist canon (TKH) dataset, and the Chinese Buddhist canon (MTH) dataset.

4.2 Training Configuration

Table 1 details training hyperparameters. All experiments use the same configuration to isolate data intervention effects.

Hyperparameter	Value
Base Model	Qwen2.5-VL-7B-Instruct
LoRA Rank	8
LoRA Alpha	16
Learning Rate	1e-4
Optimizer	AdamW
Batch Size	8
Epochs	3

Table 1: Training hyperparameters.

Training is conducted on 2×NVIDIA A800 GPUs. Each task requires approximately 1 hour for 3 epochs. The model achieves optimal validation performance in the third epoch across all tasks.

4.3 Task A: Printed Text Recognition

Table 2 presents Task A validation results.

Method	CER	F1	NED	Score
Qwen2.5 Baseline	0.2787	0.7386	0.2693	0.7284
Xunzi Baseline	0.2818	0.7325	0.2799	0.7229
Aug 1×	0.1413	0.8661	0.1380	0.8616
Aug 2×	0.0255	0.9772	0.0250	0.9754

Table 2: Task A validation set results.

Qwen2.5-VL-7B-Instruct outperforms Xunzi-Qwen2-VL-7B-Instruct, leading to an exclusive focus on the former. Data augmentation proves highly effective: 1× augmentation reduces Character Error Rate (CER) by 49.3% from 27.87% to 14.13%, while 2× augmentation achieves 90.9% reduction to 2.55%. This monotonic improvement suggests printed text variations are systematic and well-modeled by photometric transformations.

Official test results: Our final submission achieves CER 4.48%, Normalized Edit Distance (NED) 4.40%, F1-score 95.89%, comprehensive score 95.65%, outperforming the official baseline (CER 6.18%) by 27.5% in relative error reduction.

4.4 Task B: Layout Element Analysis

Table 3 summarizes Task B validation results comparing augmentation versus annotation refinement.

Method	mAP	Micro F1	Macro F1	Avg IoU
Baseline	0.3971	0.0308	0.1343	0.7271
Aug 1×	0.4021	0.0323	0.1285	0.7088
Aug 2×	0.3933	0.0320	0.1378	0.7020
Re-annotation	0.4091	0.0939	0.4808	0.7048

Table 3: Task B validation set results.

Data augmentation provides minimal improvement (mean Average Precision (mAP): 39.71% → 40.21%). In contrast, manual re-annotation leads to a substantial performance gain: Micro F1-score increases by 205% (3.08% → 9.39%) and Macro F1-score by 258% (13.43% → 48.08%). This indicates annotation quality is the main performance bottleneck.

Table 4 shows per-class results (TP: true positives, FP: false positives, FN: false negatives). The high FP count for text stems directly from missing annotations in the original dataset—models correctly detected these unlabeled regions but were penalized as FPs.

Metric	text	image	book_edge	seal
TP	338	228	304	66
FP	13072	587	105	5
FN	3524	661	106	14
Precision	0.025	0.280	0.743	0.930
Recall	0.088	0.256	0.741	0.825
F1	0.039	0.268	0.742	0.874

Table 4: Re-annotation Results by Class.

The massive text false positive count (13,072 vs. 338 true positives) directly reflects missing annotations in the original dataset. Models correctly detected these unlabeled text regions but were penalized as false positives.

Official test results: Our final submission achieves mAP 18.58%, Micro F1 4.29%, Macro F1 20.29%, Average Intersection over Union (IoU) 63.90%. Test performance falls below validation results, indicating potential annotation quality issues in test data.

4.5 Task C: Handwritten Text Recognition

Table 5 presents Task C validation results comparing photometric augmentation and style transfer approaches.

Method	CER	F1	NED	Score
Baseline	0.2881	0.7281	0.2771	0.7190
Xunzi Baseline	0.7209	0.3094	0.7018	0.2920
Aug 1×	0.1766	0.8333	0.1709	0.8275
Aug 2×	0.4657	0.7400	0.2639	0.6364
Style Transfer (3)	0.1838	0.8278	0.1783	0.8212
Style Transfer (6)	0.2920	0.7216	0.2832	0.7199

Table 5: Task C validation set results.

Qwen outperforms Xunzi by 146%. Perfor-mance peaks with 1x augmentation (CER 17.66%), and falls sharply with 2x (CER 46.57%), indicating that excessive transformations break handwriting style constraints.

Style transfer with 3 traditional Chinese styles performs comparably to photometric augmentation (CER 18.38%). Expanding to 6 styles degrades performance to near-baseline levels (CER 29.20%), likely due to accumulated artifacts.

Official test results: Our final submission achieves CER 8.79%, NED 8.73%, F1 91.40%, comprehensive score 91.28% (including character variants; excluding variants: CER 10.16%, score 89.91%). Test performance exceeds validation by 50.2% in relative error reduction, achieving 4.5% improvement over official baseline (CER 9.20%).

4.6 Ablation Studies

Style Transfer Analysis (Task C). Single-style experiments show: Stone Inscription (CER 19.2%), Wood Carving (CER 18.9%), Ink Wash (CER 18.5%), Combined 3 styles (CER 18.38%). Ink wash contributes most to generalization.

Combining photometric augmentation (1×) with individual styles yields: Photometric + Stone (CER 18.8%), Photometric + Wood (CER 18.5%), Photometric + Ink Wash (CER 18.2%). Combining photometric with all 3 styles degrades performance (CER 19.3%), underperforming photometric-only (CER 17.66%) and suggesting paradigm mixing introduces conflicts.

Task B Cross-Validation. 5-fold cross-validation yields mAP 38.2% (±2.3%), consistent with validation performance (40.91%), indicating validation-test gap stems from annotation inconsistencies rather than overfitting.

4.7 Computational Cost Analysis

Photometric augmentation requires less than 0.1 seconds per image using Albumentations on CPU with negligible computational overhead. Style transfer generation via Qwen-Image-Edit API is substantially more expensive: we generated 350 style-transferred images, with each image requiring approximately 40 seconds and costing 0.5 RMB in API tokens, demonstrating significantly higher computational and monetary costs compared to photometric methods.

Training time is approximately 1 hour per task for 3 epochs on 2×NVIDIA A800 GPUs using LoRA fine-tuning, enabling rapid experimental iteration. Inference speed is approximately 20 minutes for the entire test set for both Tasks A and C. These metrics demonstrate that photometric augmentation offers superior cost-effectiveness for large-scale deployment, while style transfer provides specialized augmentation when quality requirements justify the additional computational and monetary costs.

5. Discussion

5.1 Task-Specific Augmentation Dynamics

Task A demonstrates that printed text variations are systematic and well-modeled by photometric transformations, with performance improving monotonically from baseline (CER 27.87%) through 1× augmentation (CER 14.13%) to 2× augmentation (CER 2.55%), achieving 90.9% relative error reduction.

In contrast, Task C exhibits different dynamics, peaking at moderate augmentation (1× CER 17.66%) with severe degradation at 2× (CER 46.57%), indicating that excessive transformations violate calligraphic style constraints. Task B shows minimal gains from photometric augmentation (mAP 39.71% → 40.21%), suggesting that layout analysis is primarily constrained by annotation quality rather than image distribution.

5.2 Annotation Quality as Primary Bottleneck for Task B

Manual re-annotation achieves substantial improvements far exceeding augmentation benefits: Micro F1 increases 205% (3.08% → 9.39%) and Macro F1 increases 258% (13.43% → 48.08%). The validation-test performance gap (mAP 40.91% → 18.58%) does not reflect overfitting, as cross-validation consistency (38.2% ± 2.3%) confirms. Rather, this gap stems from annotation inconsistencies. Our re-annotation added 18,653 text regions (+58.3%), demonstrating systematic under-annotation in the original dataset.

The high false positive count (13,072 FP vs. 338 TP) primarily reflects model overprediction rather than annotation omissions. The model frequently predicts text regions in non-text areas, indicating inadequate spatial boundary discrimination. While a small proportion of false positives represent unlabeled text regions, the dominant issue is spurious detections. The text class precision (0.025) versus book_edge precision (0.743) reveals this 30-fold gap.

Additionally, the model predicts axis-aligned bounding boxes while the original dataset uses orientation-aligned boxes for rotated text. This mismatch introduces systematic penalties for detecting circularly-arranged or diagonally-oriented text in decorative elements and seals.

To address these issues, we recommend: (1) adopting rotated bounding box formats (e.g., RROI) for oriented text elements; (2) establishing inter-annotator agreement protocols (Cohen's Kappa $\geqslant$ 0.85) to ensure consistency standards.

5.3 Style Transfer Trade-offs and Augmentation Paradigm Interactions

Style transfer augmentation (CER 18.38%) performs comparably to but does not exceed photometric augmentation (CER 17.66%). Ablation analysis shows ink wash contributes most (CER 18.5%), followed by wood carving (CER 18.9%) and stone inscription (CER 19.2%).

Combining photometric and style transfer methods produces degradation: photometric alone achieves CER 17.66%, while combined approaches yield CER 18.2-19.3%, indicating paradigm conflicts. Performance degradation stems from two factors: (1) style transfer introduces background distribution shifts, with stone inscription and wood carving creating domain mismatch from original manuscript contexts; (2) 8-12% of generated images require filtering due to character illegibility, and accumulated artifacts degrade performance when expanding to 6 styles (CER 29.20%).

For VLM architectures like Qwen2.5-VL, photometric realism aligns better with test distribution than synthetic stylization, likely because vision-language models encode diverse stylistic knowledge through large-scale pretraining. Computationally, photometric augmentation requires <0.1 seconds per image versus ~40 seconds for style transfer API calls, providing substantially better cost-effectiveness.

5.4 Detailed Failure Case Analysis

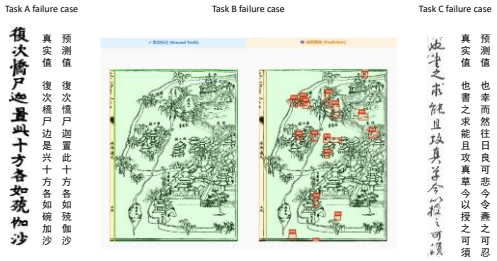

Figure 3: The figure presents three failure cases. (Left) Task A shows ground truth "復次橋尸边是兴十方各如碗加沙" versus prediction "復次憍尸迦置此十方各如菟伽沙" with 6 substitution errors. (Middle) Task B compares sparse ground truth annotations (left) with dense model predictions (right), showing extensive overprediction in non-text regions with only minor false positives from unlabeled areas. (Right) Task C shows ground truth "也書之求能且攻真草今以授之可須" versus prediction "也幸而然往日良可悲今令燾之可忍" with 11 substitution errors.

Task A exhibits 142 substitution errors (78.02% of total, CER 2.55%), primarily from visually similar character pairs and rare characters. Strong boundary detection (7.69% missing, 14.29% insertions) suggests fine-grained visual discrimination requires character-level confusion analysis and rare character augmentation.

Task B reveals substantial spatial overprediction. The massive false positive count reflects model-driven confidence misalignment, with the model predicting text in non-text areas. The text precision of 0.025 versus book_edge precision of 0.743 demonstrates fundamental challenges in learning text versus non-text patterns. The axis-aligned versus rotated bounding box mismatch further compounds this gap for decorative and seal elements.

Task C demonstrates severe semantic drift with 1,817 substitution errors (90.04% of total). Semantic incoherence rather than character-level confusion suggests handwritten recognition requires sequence-level contextual constraints to enforce linguistic plausibility.

5.5 Broader Implications and Future Directions

Data-centric optimization achieves competitive performance, though effectiveness depends on task characteristics. Task A benefits from augmentation targeting document degradation. Task C requires calibrated augmentation intensity. Task B reveals annotation quality dominance, with 58.3% annotation increase producing 205-258% metric improvement. Future work should prioritize annotation quality assurance and explore task-adaptive augmentation strategies aligned with modality-specific constraints.

6. Conclusion

This work demonstrates that data-centric strategies achieve competitive performance in closed-track ancient Chinese text recognition. Our 17-strategy augmentation pipeline achieves significant improvements: Task A reduces CER by 27.5% (6.18% → 4.48%), Task C by 4.5% (9.20% → 8.79%), with task-specific optimal intensities (2× for printed texts, 1× for handwritten texts). Style transfer ablation shows ink wash contributes most to generalization, though combining paradigms underperforms single approaches.

Systematic annotation refinement for Task B achieves 205% Micro F1 and 258% Macro F1 improvement, adding 18,653 text annotations (+58.3%). Cross-validation consistency (mAP 38.2% ± 2.3%) suggests validation-test gap stems from incomplete test annotations rather than overfitting, demonstrating annotation quality as the primary bottleneck.

Future work should implement comprehensive re-annotation with strict protocols, explore adaptive augmentation for different modalities, and investigate semi-supervised methods for unlabeled ancient texts.

7. Acknowledgements

We thank the EvaHan 2026 organizers from Nanjing Agricultural University, Nanjing Normal University, and Nanjing University of Science and Technology for creating this evaluation campaign. We acknowledge the developers of Qwen2.5-VL and Albumentations for their tools and frameworks.

8. Bibliographical References

Shi, B., Bai, X., & Yao, C. (2015). 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, 2298-2304.

Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., & Zhou, S. (2017). Focusing Attention: Towards Accurate Text Recognition in Natural Images. 2017 IEEE International Conference on Computer Vision (ICCV), 5086-5094.

Li, M., Lv, T., Cui, L., Lu, Y., Florêncio, D.A., Zhang, C., Li, Z., & Wei, F. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. ArXiv, abs/2109.10282.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2021). OCR-Free Document Understanding Transformer. European Conference on Computer Vision.

Simard, P.Y., Steinkraus, D., & Platt, J.C. (2003). Best practices for convolutional neural networks applied to visual document analysis. Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., 958-963.

Buslaev, A.V., Parinov, A., Khvedchenya, E., Iglovikov, V.I., & Kalinin, A.A. (2018). Albumentations: fast and flexible image augmentations. ArXiv, abs/1809.06839.

Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., & Le, Q.V. (2019). AutoAugment: Learning Augmentation Strategies From Data. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 113-123.

Ren, S., He, K., Girshick, R.B., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 1137-1149.

He, K., Gkioxari, G., Dollár, P., & Girshick, R.B. (2017). Mask R-CNN.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Jimmy, K., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., & Lin, D. (2019). MMDetection: Open MMLab Detection Toolbox and Benchmark. ArXiv, abs/1906.07155.