Overview of EvaHan2026: The First International Evaluation of Ancient Chinese OCR and Layout Analysis*

Dongbo Wang[1], Dongmei Zhu[1], Jieqiong Li[1], Chang Liu[1], Ruifeng Wu[1], Fan Yang[1], Yue Zhu[2,3,4], Xin Gao[1], Zhixiao Zhao[1], Xue Zhao[1], Zhongzheng Wu[5], Liu Liu[1], Bin Li[2,3,4]

[1] School of Information Management, Nanjing Agricultural University, Nanjing, China
[2] School of Chinese Language and Literature, Nanjing Normal University, China,
[3] Center for Language Big Data and Computational Humanities, Nanjing Normal University, China
[4] Lab of Artificial Intelligence and Innovative Application for Chinese Classics, Nanjing Normal University, China
[5] Artificial Intelligence and Human Languages Lab, Beijing Foreign Studies University, China
{db.wang, liuliu}@njau.edu.cn {dongmei.zhu, lijieqiong, doclc, wuruifeng, yangf, gaoxin, zhaozhixiao, zhaoxue}@stu.njau.edu.cn {wufidel}@bfsu.edu.cn {zhuyue, lib}@njnu.edu.cn

Abstract

Ancient Chinese documents are vital for historical research, necessitating high-precision character recognition and layout analysis for digitization. This paper introduces EvaHan2026, the inaugural international shared task for simultaneous optical character recognition and layout parsing of ancient texts. The evaluation framework comprehensively assesses model performance across diverse calligraphic styles and complex structures, including main body text, interlinear annotations, and illustrations. Among thirteen participating teams, four successfully completed all tasks within the closed track. Experimental results reveal that character recognition accuracy reached 97.36% on engraved texts (Test Set A) and 95.71% on handwritten texts (Test Set C) when accounting for character variants. For layout recognition in complex layouts (Test Set B), the best team achieved a peak mean Average Precision (mAP) of 59.41% and an Intersection over Union (IoU) of 76.38%. Our analysis indicates that calligraphic variability, layout density, and character variants significantly modulate system performance. Consequently, enhancing robustness within complex layouts and developing synergistic models that integrate textual and structural information remain primary challenges for intelligent interpretation of ancient writings .

Keywords: Ancient Chinese, Optical Character Recognition (OCR), Layout Analysis, Shared Task

1. Introduction

Ancient Chinese documents are pivotal for historical and philological research. Their large-scale digitization relies on high-precision optical character recognition (OCR) and layout analysis, which convert scanned images into searchable, analyzable text.

As a flagship international shared task for ancient Chinese language processing, EvaHan has built a full-stack NLP evaluation framework in collaboration with the LT4HALA/ALP workshop over four years. It started with word segmentation and POS tagging in 2022 (Li et al., 2022), expanded to machine translation in 2023 (Wang et al., 2023), added sentence segmentation and punctuation in 2024 (Li et al., 2024), and focused on named entity recognition in 2025 (Li et al., 2025). EvaHan now supports end-to-end evaluation for ancient document processing.

Massive ancient Chinese document archives have been digitized as scanned images, creating a strong demand for robust visual transcription tools (Shi et al., 2023). In this context, EvaHan 2026 for the first time centers its evaluation on OCR and layout analysis, two critical components that determine the upper bound of downstream performance. Standardized evaluation of these modules carries both academic and practical importance.

EvaHan 2026 comprises three decoupled tasks to enable fine-grained assessment: Task A (Engraved Text Recognition), Task B (Complex Layout Analysis), and Task C (Handwritten Text Recognition). Major technical challenges include pervasive character variants, print degradation, free-form handwriting, dense and interleaved layouts, paper aging, and ink bleed-through. Among these, character variants are the primary accuracy bottleneck, as they cause frequent confusions and hinder feature extraction. Accurate variant handling is essential for preserving textual fidelity and ensuring reliable downstream historical research.

Against these challenges, EvaHan 2026 aims to build a standardized, reproducible benchmark for ancient document OCR. It uses automated, objective scoring to ensure fair and consistent model comparison, thereby supporting the long-

* The official datasets and detailed guidelines for the EvaHan evaluation campaigns are available at: https://github.com/GoThereGit/EvaHan.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 244–262.
11 May 2026.

term preservation and scholarly use of cultural heritage.

2. Related Work

2.1 Deep Learning-Based OCR Methodologies

Deep learning has fundamentally advanced OCR systems. Early gradient‑based learning and Graph Transformer Networks (GTN) laid the foundation for end‑to-end trainable document recognition (LeCun et al., 1998). The Convolutional Recurrent Neural Network (CRNN) integrated CNN and RNN features to enable direct sequence recognition without explicit character segmentation, becoming a de facto standard for text recognition (Shi et al., 2016). More recently, OCR‑free frameworks such as Donut directly transformed document images into structured information, alleviating error propagation and excessive computation in traditional pipelines (Kim et al., 2021). These technical breakthroughs provide a solid basis for visual analysis and information extraction from ancient documents.

2.2 Visual Processing for Historical Documents

Research on historical document visual processing has advanced rapidly across pipelines, benchmarks, layout analysis, and multi-task modeling. OCR4all delivered a semi‑automated workflow for 15th–19th century engraved books, achieving low CER but still requiring manual proofreading for complex layouts (Reul et al., 2019). M⁵HisDoc constructed a large‑scale benchmark with eight thousand images and a multi‑attribute difficulty definition system, yet performance degrades on seriously distorted samples (Shi et al., 2023). HierText built hierarchical layout parsing standards across word, line, and paragraph levels (Long et al., 2022). Xu et al. (2018) designed a multi‑task FCN for pixel‑level segmentation of pre‑modern Chinese handwritten manuscripts. AncientDoc pioneered VLM evaluation for ancient documents, covering tasks from page‑level OCR to knowledge reasoning (Yu et al., 2025). Despite these advances, a unified, international benchmark for simultaneous OCR and layout analysis remains lacking.

2.3 Shared Tasks and Evaluation Platforms

Specialized competitions and platforms have accelerated technological development. The Guangdong‑Hong Kong‑Macao Greater Bay Area (Huangpu) International Algorithm Competition has launched a dedicated track for Ancient Document Image Recognition and Analysis, with supporting resources available at https://github.com/SCUT-DLVCLab/MCS-Bench. However, existing evaluations rarely support decoupled yet unified testing of Engraved Text, Handwritten Text, and Complex Layout Analysis. Few systematically address character variants or compare model performance under closed versus open settings. Building on the established EvaHan series, EvaHan 2026 fills these gaps with a holistic, standardized, and internationally accessible evaluation framework for ancient Chinese document visual processing.

3. Task Description

To systematically address the technical bottlenecks inherent in digitizing heterogeneous ancient Chinese texts, EvaHan 2026 structures its evaluation into three distinct sub-tasks based on visual characteristics and target objectives (summarized in Table 1). This decoupled task design is specifically formulated to quantify model performance across the distinct challenges of block-engraved character recognition, complex layout parsing, and manuscript transcription.

Task A evaluates the model's recognition capability on engraved classical texts across different manuscript styles. Engraved classical texts not only involve character evolution and complex layout arrangements (such as marginal annotations and double-line footnotes), but also commonly suffer from degradation phenomena like missing pages, ink stains, and damaged or missing characters, making their recognition significantly more challenging than standardized modern texts. This task employs Character Error Rate (CER) as the primary evaluation metric, supplemented by F1-score and Normalized Edit Distance (NED). This approach aims to comprehensively evaluate models across multiple dimensions, including recognition accuracy, sequence integrity, and transcription cost.

Task B focuses on Document Layout Analysis, aiming to achieve automated extraction of key semantic elements within complex classical text layouts. The identified objects encompass four core components: text blocks, illustrations, book margins, and seals. The evaluation employs Mean Average Precision (mAP) as the primary metric, while incorporating Intersection over Union (IoU) and F1 score to scientifically quantify the model's overall performance in regional localization accuracy and multi-class recognition tasks.

Task C targets recognizing handwritten ancient texts. Manuscripts exhibit highly individualized characteristics and are rife with random disturbances such as hasty connected strokes, colloquial/variant characters, missing strokes, and later alterations, creating significant technical barriers. This task comprehensively employs CER, F1 score, and NED metrics to overcome algorithmic bottlenecks in automatic transcription of handwritten text, providing foundational

technical support for deep mining of rare unique copies and manuscripts, as well as knowledge graph construction.

4. Dataset

4.1 Data Sources

The EvaHan2026 evaluation task utilizes datasets curated from highly representative classical Chinese texts, aiming to comprehensively cover diverse textual formats and layout characteristics. Both Datasets A and B are sourced from the Siku Quanshu, a monumental collection of immense scholarly value. Dataset A features extensive sampling from the Four Divisions: Confucian Classics, History, Philosophy, and Literature, encompassing diverse fields such as philosophy, history, and literature, and providing abundant engraved text samples. Concurrently, Dataset B is derived from complex pages in these classical texts that feature a mix of text and images.

Dataset C focuses on the more challenging task of handwriting text recognition, featuring sources with significant heterogeneity and cross-genre distribution. These materials can be broadly categorized into four main types. First, historical and governmental archives, such as the Chronicle of Foreign Affairs Arrangements, document Qing dynasty diplomatic affairs. Second, private literary collections and poetry anthologies—exemplified by the Poems of the Escaping Void Master and Collected Works of Master Renjie—showcase highly individualized writing styles. Third, calligraphic manuscripts and ink tracings, such as the Bu Shang Reading Manuscript, represent highly artistic and challenging handwriting. Finally, core Confucian and religious literature encompasses the Thirteen Classics and multiple volumes of Chinese Buddhist Scriptures. This corpus, spanning engraved and handwritten materials as well as official and folk sources, ensures the evaluation task can assess the robustness of OCR algorithms across multiple dimensions in complex real-world scenarios.

4.2 Data Annotation

To ensure the model can learn deep semantic and visual features, this evaluation conducted targeted and manually refined annotation on three sub-datasets. Examples from the three datasets are shown in Table 1.

The annotation team performed character-level annotation and variant character processing for Datasets A and C, achieving bounding box localization and text transcription for images. Addressing the prevalent issue of variant characters in ancient texts, the annotation strategy balanced “character form priority” with “semantic consideration.” The annotation guidelines also standardized and consolidated exceptionally rare variant characters. For example, when the character 𣍐 appeared with its bottom component written as 壬 instead of 王, it was consistently annotated as 呈 in accordance with our unified transcription specifications.

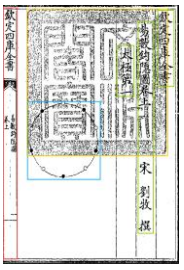
Dataset	Example	Text
Dataset A		{"image_path": "Dataset_A/a_016.jpg", "text": "管民官管匠人打捕諸頭目及諸軍馬使臣人等宣聖廟國家歲"},
Dataset B		{"image_path": "Dataset_B/b_0002.jpg","regions": [{"label": "book_edge", "text": "" "points": [[2, 12], [42, 12], [42, 776], [2, 776]]}, {"label": "image"……
Dataset C		{"image_path": "Dataset_C/c_015.jpg","text": "臣奉旨知道了欽此"},

Table 1: Dataset Sample

For complex layout annotation, Dataset B employs a rectangular region annotation strategy to delineate boundaries for text, images, book margins, and seals. Addressing the unique layout complexity of ancient texts, this study focuses on analyzing the proportion of text-image interleaved data. Approximately 27.46% of the dataset features highly integrated image regions with main text and flexible layouts. Such high-complexity annotations require models to possess not only fundamental object detection capabilities but also the ability to handle hierarchical layout relationships.

For each subtask, we have set up 5,000 images for training and 200 images for test. During the training phase, participants can only access the annotated training set for parameter optimization and internal validation (as detailed in Table 2).

Dataset	Images	Char tokens	Char types
Dataset A training	5,000	83,832	3,241
Dataset A test	200	4,589	1,415
Dataset B training	5,000	-	-
Dataset_B test	200	-	-
Dataset C training	5,000	69,183	4,729
Dataset C test	200	4,728	681
Total	-	162,332	5,551

Table 2: Dataset Statistics

5. Evaluation Metrics

This evaluation encompasses three core tasks: Task A (Engraved Text Recognition), Task B (Complex Layout Analysis), and Task C (Handwritten Text Recognition). To comprehensively and objectively assess model performance, we adopt a suite of well-established metrics tailored to text recognition and object detection tasks.

For text recognition tasks (Tasks A and C), we use five complementary metrics: CER as the primary core metric, alongside Precision, Recall, F1-Score, and NED to capture recognition accuracy, completeness, and overall edit distance efficiency. For layout analysis (Task B), we employ mean Average Precision (mAP@[.5:.95]) as the primary detection metric, supplemented by Micro-average F1, Macro-average F1, and Average Matching IoU to evaluate category-level detection performance, class balance, and pixel-level localization precision, respectively. Detailed definitions of auxiliary metrics and the final comprehensive ranking score can be found in the Appendix.

5.1 Text Recognition Metrics (Task A and Task C)

In the text recognition phase of Tasks A and C, we adopted a character-level evaluation approach. This system primarily quantifies recognition errors by calculating the Edit Distance between the model's predicted text and the ground truth labels. The specific evaluation metrics include the following dimensions:

(1) Character Error Rate (CER)

CER is the core metric of this evaluation, defined as the ratio of the Edit Distance to the length of the reference text:

$$\mathrm{CER}=\frac{\mathrm{Edit\backslash Distance(ref,hyp)}}{|\mathrm{ref}|} \qquad (5\text{-}1)$$

where *ref* is the reference text (ground truth), h*yp* is the predicted text (OCR output), and |*ref*| represents the number of characters in the reference text.

The calculation of the Edit Distance includes three basic operations:

$$\mathrm{Edit(ref,hyp)}=\mathrm{N_{del}}+\mathrm{N_{ins}}+\mathrm{N_{sub}} \qquad (5\text{-}2)$$

Deletion: N_{del}, the number of characters present in the reference text but missing in the predicted text.

Insertion: N_{ins}, the number of characters present in the predicted text but not in the reference text.

Substitution: N_{sub} , the number of characters present in both but different.

(2) Variant Character Matching

Considering the existence of a large number of variant characters in ancient texts, Tasks A and C provide two evaluation modes. The standard mode counts a prediction as correct only when the predicted character matches the reference character exactly; the variant character mode is based on a predefined variant character mapping table, where a predicted character is also counted as correct if it is a visually similar variant of the reference character. Under the variant character mode, the CER calculation will automatically match equivalent characters within the edit distance algorithm. The equivalence relation sim is defined as:

$$c_1 \sim c_2 \Leftrightarrow c_1=c_2 \vee (c_1,c_2)\in V \qquad (5\text{-}3)$$

where *V* is the set of variant character mappings. During the edit distance calculation, if *ref*[*i*]~h*yp*[*j*], it is considered a successful match. Specifically, the evaluation code uses a dynamic programming algorithm to calculate the edit distance, introducing variant character judgment during the character alignment phase: when a mapping relationship exists in *V* between the reference character and the predicted character, the two are considered equivalent, and no insertion, deletion, or substitution operations are required. Regarding the selection of variant characters, researchers manually filtered high-frequency variant character pairs from the test set files of datasets A and C to build the variant character mapping table used for the evaluation calculation.

5.2 Text Recognition Metrics (Task B)

Task B adopts object detection evaluation standards to assess the model's ability to localize and classify layout elements (text blocks, illustrations, book margins, and seals). The primary metric used is the Mean Average Precision (mAP).

(1) Mean Average Precision (mAP)

We adopt the mAP@[0.5:0.95] of the COCO evaluation standard (Lin et al., 2014). Specifically, we sequentially calculate the AP at IoU thresholds ranging from 0.5 to 0.95 (with a step size of 0.05), and finally calculate the mean. During this process, the AP calculation at a single threshold refers to the classic 11-point interpolation method, ensuring evaluation smoothness and objectivity by sampling on the precision-recall curve. That is, calculating the mean of average precisions when the IoU threshold goes from 0.5 to 0.95 (step size 0.05):

$$\mathrm{mAP_{[0.5:0.95]}}=\frac{1}{10}\sum_{t\in\{0.50,0.55,\ldots,0.95\}}\mathrm{AP}\,(t) \qquad (5\text{-}4)$$

The Average Precision (*AP*) at each IoU threshold *t* is calculated via the 11-point interpolation method on the precision-recall curve:

$$\mathrm{AP}(t)=\frac{1}{11}\sum_{r\in\{0.0,0.1,\ldots,1.0\}}\max_{r'\geq r}\mathrm{P}\,(r') \qquad (5\text{-}5)$$

where $P(r')$ is the precision at recall r'.

(2) Intersection over Union (IoU)

For any predicted bounding box B_p and ground truth bounding boxB_g, IoU is defined as:

$$\mathrm{IoU}(\mathrm{B_p},\mathrm{B_g})=\frac{\mathrm{Area}(\mathrm{B_p}\cap\mathrm{B_g})}{\mathrm{Area}(\mathrm{B_p}\cup\mathrm{B_g})} \quad (5\text{-}6)$$

The matching rule uses a greedy algorithm: labels are grouped for each image, the IoU matrix between predicted boxes and ground truth boxes is calculated, and the predicted-ground truth box pair with the maximum IoU is preferentially matched (IoU ≥0.5 is considered a valid match).

5.3 Evaluation Modalities

This evaluation sets up two participation modalities: the Closed Modality and the Open Modality, to adapt to the needs of different technical routes and practical application scenarios.

(1) **Closed Modality**

Under the Closed Modality, participants are strictly limited to using the officially provided pre-trained models and the fixed training set data. This modality aims to evaluate the generalization ability of models under restricted resource conditions, simulating the real-world constraints of limited annotated data and computing resources in actual ancient document digitization scenarios. The authorized pre-trained models allowed in this modality include two VLMs, Qwen2.5-7B-VL and Xunzi-Qwen2-7B-VL, as well as the YOLO series object detection models. Qwen2.5-7B-VL is a multi-modal large language model launched by Alibaba (Bai et al., 2023), possessing strong image understanding and text generation capabilities, and performing excellently in general multi-modal data processing tasks; Xunzi-Qwen2-7B-VL (Wang et al., 2024) is a specialized model optimized for ancient text recognition scenarios, featuring targeted advantages in ancient character recognition and layout understanding; the YOLO series models are classic object detection architectures, widely applied in layout element localization tasks. By restricting model selection and data sources, the closed modality facilitates a fair benchmarking of the effectiveness of different methods, and its evaluation results better reflect the generalization ability of models under data-constrained conditions.

(2) Open Modality

Under the Open Modality, the evaluation does not restrict the use of models and training data, encouraging participants to explore state-of-the-art technical methods. Participants are free to use various multi-modal large models including GPT-4o-mini, LLaVA (Liu et al., 2024), Deepseek-OCR (Lu et al., 2024), etc. They can also utilize publicly available ancient book datasets and synthetic data for training, and may introduce external knowledge bases such as variant character dictionaries and ancient character sets to enhance model performance. This modality aims to push the technical boundaries of ancient book OCR and layout analysis, evaluating the optimal performance of current methods under unconstrained conditions. It is expected that the open modality will achieve higher evaluation scores, but it may also be accompanied by higher computational costs. By comparing the evaluation results of the two modalities, we hope to comprehensively reveal the current technical level of ancient book OCR and layout analysis from the dual dimensions of algorithmic effectiveness and technical upper limits, providing a reference basis for actual deployment.

6. Participants and Results

6.1 Participating Teams

This section statistically analyses the institutional affiliations, participation scale, and model submission features of teams in the EvaHan 2026 competition, based on official application and final submission data. Initially, 41 teams registered, with core data drawn from the organizing committee's registration forms and technical review database. Ultimately, 13 teams successfully submitted their final evaluation results across the three core tasks (Task A for Engraved OCR, Task B for Layout Analysis, and Task C for Handwritten OCR). Among them, 4 teams participated in both the closed and open modalities, while the remaining teams focused solely on the closed track. It is worth emphasizing that, with the singular exception of SL-ISTIS–which completed the model evaluation but failed to provide the corresponding documentation–all other 12 participating teams successfully submitted their required final papers. Participating entities consist primarily of universities and research teams, with the detailed information of each team's model and method submissions presented in Table 3.

6.2 Results

The performance evaluation of this competition was structured around three core tasks: Task A (Engraved Text Recognition), Task B (Complex Layout Analysis), and Task C (Handwritten Text Recognition). To comprehensively assess the systems, we established two evaluation tracks (closed and open modalities) and applied two distinct character matching rules (inclusive and exclusive of variant characters). Owing to the large number of participating teams in this EvaHan edition and the page limit of the paper, only the top-five results under different tracks and tasks are presented for the evaluation results of each participating team. Complete rankings are available in Appendix A.

Visualizing the performance metrics across modalities reveals critical technical insights as

shown in Figure 1. In the closed modality, TJU's specialized architectural design such as the integration of HistLayout-DETR achieved state-of-the-art results in the OCR tasks (Tasks A and C). This demonstrates that task-specific structural innovations can effectively compensate for strict resource constraints. Conversely, WHU-SAI relied on comprehensive multi-stage optimization (LoRA SFT, DPO, GRPO), maintaining dominant positions across the open modality. However, as indicated by the tight convergence of the open and closed trend lines for top-performing systems, leveraging unrestricted external datasets and larger model ensembles yielded only marginal gains in peak performance—and even a slight degradation in Task B for the leading team. Notably, while the closed modality exhibits severe performance drop-offs (indicated by the sharp upward spikes in CER) among lower-ranked teams, the open track maintains a remarkably flat trajectory. This suggests that scaling data and model size primarily functions to elevate the performance floor and ensure system robustness, rather than significantly advancing the absolute technical ceiling. Ultimately, these patterns highlight a diminishing return on simply scaling general multimodal models without introducing corresponding, domain-specific algorithmic innovations.

ID	Team	Affiliation	Close/ Open	Core Models	Key Methods & Data
1	BNU	Beijing Normal University	2/0	Xunzi-VL, Qwen-VL, Doubao-1.5-pro (Vision)	LoRA, linguistic rule post-processing; No external data
2	ENC-PSL	École nationale des chartes, PSL University	2/2	CRNN, PaddleOCR, YOLO series, RT-DETR, Qwen-VL series, GPT-4o, Claude, Gemini	UltraGlyph pipeline, ICL, manual re-annotation; External data used
3	EPHE-PSL	École pratique des hautes études, PSL University	1/1	ANANDASKY VLM, ViT, Qwen3-0.6B, YOLO26	Global attention encoder, line-level transcription, full fine-tuning; External data used
4	FDU	Fudan University	1/0	Xunzi-VL, DocLayout-YOLOv10	SFT, DoRA, DPO, multi-round inference; No external data
5	NEFU	Northeast Forestry University	1/0	Qwen-VL, YOLOv8/11	LoRA, dual-stream YOLO ensemble, VLM semantic filter; No external data
6	NJU	Nanjing University	1/0	Qwen-VL	LoRA, QLoRA, adaptive prompt engineering; No external data
7	QLNU	Qilu Normal University	1/0	Qwen-VL	LoRA, data augmentation, annotation refinement; No external data
8	RUC-MIDU	Renmin University of China; Midu Technology	2/2	Qwen-VL, Xunzi-VL, DocLayout-YOLO	LoRA, background embedding augmentation; No external data
9	SL-ISTIS	Shanghai Library (Institute of Scientific and Technical Information of Shanghai)	1/0	-	No valid paper submitted
10	SYSU	Sun Yat-sen University	2/0	Qwen-VL	SFT, LoRA, full-parameter FT, GRPO, curriculum learning; No external data
11	TJU	Tongji University	2/0	Qwen-VL, Xunzi-VL, HistLayout-DETR, ResNet-50	LoRA, coordinate normalization, curriculum learning; No external data
12	WHU	Wuhan University	2/0	Qwen-VL	LoRA, multi-task joint training; No external data
13	WHU-SAI	Wuhan University (School of Artificial Intelligence)	2/2	Qwen-VL	SFT, LoRA, DPO, GRPO, curriculum learning; External data used

Table 3: Overview of participating teams by modality ("Close/Open" denotes the number of evaluation result submissions by participating teams in the closed and open tracks, respectively.)

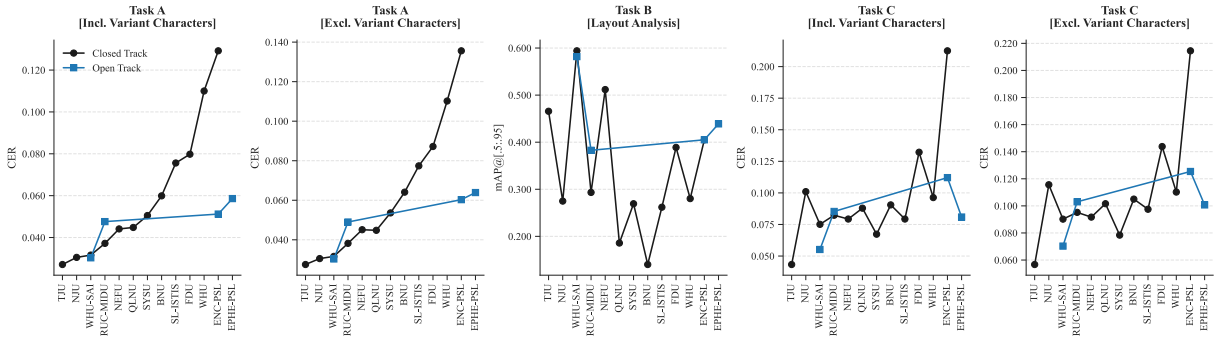


Figure 1: Performance by Task and Modality

6.2.1 Evaluation Results under Closed Modality

Table 4 and Table 5 summarize the top-five results under variant-included and variant-excluded rules.

ID	Team	Task A (inclusive of variant chars)			
		CER	NED	F1	Comp
1	TJU	2.71%	2.70%	97.54%	97.36%
2	NJU	3.05%	3.02%	97.19%	97.03%
3	WHU--SAI	3.16%	3.14%	97.02%	96.90%
4	RUC-MIDU	3.72%	3.69%	96.63%	96.39%
5	NEFU	4.41%	4.35%	95.89%	95.69%
ID	**Team**	**Task A (with variant chars excluded)**			
		CER	**NED**	**F1**	**Comp**
1	TJU	2.75%	2.74%	97.50%	97.32%
2	NJU	3.05%	3.02%	97.19%	97.03%
3	WHU--SAI	3.16%	3.14%	97.02%	96.90%
4	RUC-MIDU	3.82%	3.79%	96.53%	96.29%
5	QLNU	4.48%	4.40%	95.89%	95.65%

Table 4: Top5 results of Task A closed

TJU achieved state-of-the-art performance across both engraved and handwritten recognition tasks. For Task A, TJU attained a Comprehensive Score of 0.9736 (variant-included) with a CER of 0.0271, leveraging an innovative integration of Qwen2.5-VL with HistLayout-DETR (Carion et al., 2020). This layout-aware framework effectively handled complex typographical structures inherent in block-engraved text. For Task C, TJU secured first place with a score of 0.9571, outperforming the second-place team by 2.4 percentage points. NJU and WHU-SAI followed closely in Task A, utilizing base Qwen2.5-VL-7B and multi-stage optimization strategies such as Lora SFT (Hu et al., 2021), DPO (Rafailov et al., 2024), and GRPO (Shao et al., 2024), respectively.

ID	Team	Task C (with variant chars included)			
		CER	NED	F1	Comp
1	TJU	4.33%	4.33%	95.80%	95.71%
2	SYSU	6.73%	6.72%	93.39%	93.31%
3	WHU-SAI	7.51%	7.50%	92.67%	92.55%
4	NEFU	7.93%	7.91%	92.22%	92.12%
5	SL-ISTIS	7.93%	7.93%	92.16%	92.10%
ID	**Team**	**Task C (exclusive of variant chars)**			
		CER	**NED**	**F1**	**Comp**
1	TJU	5.67%	5.66%	94.46%	94.37%
2	SYSU	7.84%	7.83%	92.28%	92.20%
3	WHU-SAI	9.02%	9.01%	91.16%	91.04%
4	NEFU	9.17%	9.15%	90.98%	90.88%
5	RUC-MIDU	9.52%	9.51%	90.68%	90.54%

Table 5: Top5 results of Task C closed

For layout analysis (Task B), metrics included mAP@[.5:.95], Micro/Macro F1-scores, and Avg Match IoU. Table 6 presents the top-five results.

ID	Team	Task B			
		mAP@ [.5:.95]	Micro F1	Macro F1	Avg Match IoU
1	WHU-SAI	59.41%	83.42%	86.86%	76.38%
2	NEFU	51.18%	75.53%	78.79%	77.40%
3	TJU	46.57%	57.42%	71.34%	74.30%
4	ENC-PSL	40.51%	70.83%	67.75%	74.03%
5	FDU	38.89%	64.01%	63.18%	81.19%

Table 6: Top5 results of Task B closed

WHU-SAI achieved optimal performance with an mAP of 0.5941, significantly surpassing the

second-place team (NEFU, 0.5118). Their success stemmed from a sophisticated multi-stage training strategy combining SFT, DPO, and GRPO on Qwen2.5-VL-7B, augmented with LoRA for efficient parameter tuning. NEFU employed a dual-stream architecture combining YOLO ensembles with Qwen2.5-VL semantic filtering. Notably, FDU achieved the highest Avg Match IoU (0.8119), indicating superior bounding box localization accuracy despite lower classification performance.

6.2.2 Evaluation Results under Open Modality

The open modality imposes no additional restrictions on the pre-trained models, external data, and training strategies used by the participating systems, which can better test the upper performance limit of current mainstream multimodal large models and related technologies in ancient document visual processing tasks. A total of four teams submitted valid results for the open track of this competition, and the core performance of each system is as follows.

ID	Team	Task B			
		mAP@ [.5:.95]	Micro F1	Macro F1	Avg Match IoU
1	WHU-SAI	58.18%	82.78 %	86.45 %	75.72%
2	EPHE-PSL	43.89%	72.32 %	66.57 %	81.14%
3	ENC-PSL	40.51%	70.83 %	67.75 %	74.03%
4	RUC-MIDU	38.28%	66.73 %	59.83 %	80.41%

Table 8: Evaluation results of Task B open

The evaluation results of the character recognition tasks under the open modality are shown in Table 7.

WHU-SAI demonstrated consistent superiority across all open modality tasks. For Task A, their system attained a comprehensive score of 0.9703 (CER: 0.0303), approaching the closed modality optimum (0.9736). This marginal gap (0.33%) suggests their closed-track optimization strategy was already near the technical frontier. For Task C, WHU-SAI maintained leadership with scores of 0.9456 (variant-included) and 0.9305 (variant-excluded), utilizing a multi-stage paradigm (SFT, LoRA, DPO, GRPO) effective for handling calligraphic variability.

The evaluation results of the layout analysis task under the open modality are shown in Table 8. In layout analysis (Task B), WHU-SAI achieved an mAP of 0.5818, only 2.1% lower than their closed modality result. EPHE-PSL secured second place (mAP: 0.4389) by leveraging a multiscript pretraining corpus and 4.04 million annotated lines of historical documents. The marginal performance gains overall suggest that carefully curated training strategies tailored for the closed track may be more pivotal than simply scaling data and model parameters.

ID	Team	Task A (with variant chars included)			
		CER	NED	F1	Comprehensive
1	WHU-SAI	3.03 %	3.01%	97.15 %	97.03%
2	RUC-MIDU	4.76 %	4.67%	95.60 %	95.37%
3	ENC-PSL	5.12 %	5.08%	95.32 %	95.02%
4	EPHE-PSL	5.86 %	5.64%	94.79 %	94.38%
ID	**Team**	**Task A (with variant chars excluded)**			
		CER	**NED**	**F1**	**Comprehensive**
1	WHU-SAI	3.03 %	3.01%	97.15 %	97.03%
2	RUC-MIDU	4.90 %	4.81%	95.46 %	95.23%
3	EPHE-PSL	6.03 %	5.99%	94.41 %	94.11%
4	ENC-PSL	6.39 %	6.17%	94.26 %	93.85%
ID	**Team**	**Task C (with variant chars included)**			
		CER	**NED**	**F1**	**Comprehensive**
1	WHU-SAI	5.52 %	5.49%	94.71 %	94.56%
2	EPHE-PSL	8.09 %	8.09%	92.00 %	91.94%
3	RUC-MIDU	8.53 %	8.49%	91.68 %	91.54%
4	ENC-PSL	11.21 %	10.90%	89.96 %	89.20%
ID	**Team**	**Task C (with variant chars excluded)**			
		CER	**NED**	**F1**	**Comprehensive**
1	WHU-SAI	7.03 %	7.00%	93.20 %	93.05%
2	EPHE-PSL	10.07 %	10.07%	90.02 %	89.96%
3	RUC-MIDU	10.30 %	10.26%	89.91 %	89.77%
4	ENC-PSL	12.54 %	12.22%	88.63 %	87.88%

Table 7: Top5 results of Tasks A and C open

7. Conclusion

The evaluation task focused on character recognition and structural parsing within complex classical documents that are characterized by heterogeneous calligraphic styles and intricate page formats. Despite formidable challenges

including orthographic variants, layout interference, and interleaved text-image arrangements, the majority of participating teams demonstrated notable robustness, and achieved consistent and reliable performance across various subtasks. The results indicate that performance on well-structured engraved texts generally surpasses that achieved on handwritten manuscripts. Although the latter presents greater challenges, the majority of participating systems substantially exceeded the baseline models. This demonstrates the robust efficacy of contemporary methodologies in addressing the complexities of ancient document corpora. Methodologically, most teams developed task-specific models for distinct subtasks, resulting in significant performance gains in both character recognition and layout region detection. However, a unified architecture capable of concurrently processing diverse calligraphic styles and complex structures remains elusive, as does the synergistic modeling of textual content and layout semantics. Furthermore, comparative analysis reveals that zero-shot inference relying solely on general-purpose large VLMs underperforms in specialized ancient document contexts, underscoring the necessity of domain-specific instruction fine-tuning and layout-aware architectural adaptations.

Future research should focus on two primary directions: advancing lightweight specialized models to capture the nuances of ancient character morphology, and integrating LLMs into hybrid frameworks to balance stability and generalization. We expect future work to achieve high-precision, integrated recognition across heterogeneous writing styles and layouts. Beyond the scope of this task, the field is evolving toward more complex challenges, including structured text generation, layout-to-text alignment, and multimodal understanding of ancient Chinese documents.

8. Acknowledgments

This work was supported by the Major Project of the National Social Science Fund of China (Grant No. 21&ZD331) and General Program of the Ministry of Education of China Humanities and Social Sciences Fund (Grant No. 24A10319028).

9. Ethical Considerations

The datasets utilized in this evaluation are sourced exclusively from historical Chinese documents. Any views, ideologies, or cultural norms reflected in the textual content are inherent to the original ancient corpora and do not represent the perspectives, values, or endorsements of the authors or the organizing committee.

10. Bibliographical References

Li, B., Y. Yuan, J. Lu, M. Feng, C. Xu, W. Qu, and D. Wang. 2022. The First International Ancient Chinese Word Segmentation and POS Tagging Bakeoff: Overview of the EvaHan 2022 Evaluation Campaign. Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, pages 135–140.

Wang, D., L. Lin, Z. Zhao, W. Ye, K. Meng, W. Sun, L. Zhao, X. Zhao, S. Shen, W. Zhang, and B. Li. 2023. EvaHan2023: Overview of the First International Ancient Chinese Translation Bakeoff. Proceedings of ALT2023: First Workshop on Ancient Language Translation, pages 1–14.

Li, B., B. Chang, Z. Xu, M. Feng, C. Xu, W. Qu, S. Shen, and D. Wang. 2024. Overview of EvaHan2024: The First International Evaluation on Ancient Chinese Sentence Segmentation and Punctuation. Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2024) @ LREC-COLING-2024, pages 229–236.

Li, B., B. Chang, R. Liu, X. Zhao, S. Shen, L. Liu, Y. Zhu, Z. Xu, W. Qu, and D. Wang. 2025. Overview of EvaHan2025: The First International Evaluation on Ancient Chinese Named Entity Recognition. Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 97–105.

Bai, J. Z., S. Bai, S. S. Yang, S. J. Wang, S. N. Tan, P. Wang, J. Y. Lin, C. Zhou, and J. R. Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966.

Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. 2020. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), pages 213–229, Cham. Springer.

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

Kim, G., T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park. 2021. OCR-free document understanding transformer. In Proceedings of the European Conference on Computer Vision (ECCV).

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324.

Li, B., B. L. Chang, R. L. Liu, X. Zhao, S. Shen, L. H. Liu, Y. Zhu, Z. X. Xu, W. G. Qu, and D. B. Wang. 2025. Overview of EvaHan2025: The

first international evaluation on ancient Chinese named entity recognition. In Proceedings of the Second Workshop on Ancient Language Processing, pages 156–164, Laguna. Association for Computational Linguistics.

Li, B., B. L. Chang, Z. X. Xu, M. X. Feng, C. Xu, W. G. Qu, S. Shen, and D. B. Wang. 2024. 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, Torino, Italia. ELRA and ICCL.

Li, B., Y. G. Yuan, J. Y. Lu, M. X. Feng, C. Xu, W. G. Qu, and D. B. Wang. 2022. The first international ancient Chinese word segmentation and POS tagging bakeoff: Overview of the EvaHan 2022 evaluation campaign. In Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, pages 135–140, Marseille, France. European Language Resources Association.

Lin, T. Y., M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. 2014. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), pages 740–755, Cham. Springer.

Liu, H., C. Li, Q. Wu, and Y. J. Lee. 2024. Visual instruction tuning. In Advances in Neural Information Processing Systems (NeurIPS), 36.

Long, S. B., S. Y. Qin, D. Panteleev, A. Bissacco, Y. Fujii, and M. Raptis. 2022. Towards end-to-end unified scene text detection and layout analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1039–1049.

Lu, H., W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. X. Sun, T. Z. Ren, Z. S. Li, H. Yang, Y. F. Sun, C. Q. Deng, H. W. Xu, Z. D. Xie, and C. Ruan. 2024. DeepSeek-VL: Towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525.

Rafailov, R., A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems (NeurIPS), 36.

Reul, C., D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, and F. Puppe. 2019. OCR4all - an open-source tool providing a (semi-)automatic OCR workflow for historical printings. Applied Sciences, 9(22):4853.

Shao, Z. H., P. Y. Wang, Q. H. Zhu, R. X. Xu, J. X. Song, X. Bi, H. W. Zhang, M. C. Zhang, Y. K. Li, Y. Wu, and D. Y. Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.

Shen, Z., R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, and W. Li. 2021. LayoutParser: A unified toolkit for deep learning based document image analysis. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 131–146. Springer International Publishing.

Shi, B., X. Bai, and C. Yao. 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.

Shi, Y., C. Liu, D. Peng, C. Jian, J. Huang, and L. Jin. 2023. M5HisDoc: A large-scale multi-style Chinese historical document analysis benchmark. In Advances in Neural Information Processing Systems (NeurIPS), 36.

Wang, D. B., L. T. Lin, Z. X. Zhao, W. H. Ye, K. Meng, W. L. Sun, L. Z. Zhao, X. Zhao, S. Shen, W. Zhang, and B. Li. 2023. EvaHan2023: Overview of the first international ancient Chinese translation bakeoff. In Proceedings of the Ancient Language Translation Workshop (ALT2023), pages 1–14, Macao SAR, China. Asia-Pacific Association for Machine Translation.

Wang, T., Y. Wu, R. Rubino, and C. Baumgartner. 2019. Radical aggregation network for few-shot offline handwritten Chinese character recognition. Pattern Recognition Letters, 125:821–827.

Wang, X., Y. Wang, B. Li, and D. Wang. 2024. Xunzi: A Chinese ancient text large language model. In Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 11984–11993.

Xu, Y., F. Yin, Z. X. Zhang, and C. L. Liu. 2018. Multi-task layout analysis for historical handwritten documents using fully convolutional networks. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18), pages 1057–1063. AAAI Press.

Yu, H. Y., Y. C. Wu, F. Shi, L. Liao, J. H. Lu, X. D. Ge, H. Wang, M. H. Zhuo, X. C. Wu, X. Fei, H. Feng, G. Z. Tang, A. L. Wang, H. S. Zhu, Y. F. He, Q. H. Liang, L. Y. Meng, C. Feng, C. Huang, J. Q. Tang, and B. Li. 2025. Benchmarking vision-language models on Chinese ancient documents: From OCR to knowledge reasoning. arXiv preprint arXiv:2509.09731.

Appendix A: Scores of Participating Teams

A.1 Scores of Participating Teams under Closed Modality

For comprehensive reference, the complete ranking results for all participating teams across all tasks and modalities are provided below. These tables include all valid submissions received during the evaluation period, extending beyond the top-five rankings presented in the main text.

ID	Team	Task A (with variant characters included)			
		CER	NED	F1	Comprehensive
1	TJU	2.71%	2.70%	97.54%	97.36%
2	NJU	3.05%	3.02%	97.19%	97.03%
3	WHU--SAI	3.16%	3.14%	97.02%	96.90%
4	RUC-MIDU	3.72%	3.69%	96.63%	96.39%
5	NEFU	4.41%	4.35%	95.89%	95.69%
6	QLNU	4.48%	4.40%	95.89%	95.65%
7	SYSU	5.05%	5.02%	95.36%	95.08%
8	BNU	5.99%	5.95%	94.53%	94.17%
9	SL-ISTIS	7.56%	7.44%	93.06%	92.65%
10	FDU	7.98%	7.95%	92.70%	92.23%
11	WHU	11.00%	10.89%	89.36%	89.13%
12	ENC-PSL	12.92%	12.89%	87.73%	87.28%
ID	Team	Task A (with variant characters excluded)			
		CER	NED	F1	Comprehensive
1	TJU	2.75%	2.74%	97.50%	97.32%
2	NJU	3.05%	3.02%	97.19%	97.03%
3	WHU--SAI	3.16%	3.14%	97.02%	96.90%
4	RUC-MIDU	3.82%	3.79%	96.53%	96.29%
5	QLNU	4.48%	4.40%	95.89%	95.65%
6	NEFU	4.51%	4.46%	95.79%	95.59%
7	SYSU	5.36%	5.32%	95.05%	94.77%
8	BNU	6.41%	6.37%	94.11%	93.75%
9	SL-ISTIS	7.74%	7.62%	92.88%	92.47%
10	FDU	8.72%	8.69%	91.95%	91.49%
11	WHU	11.02%	10.91%	89.34%	89.11%
12	ENC-PSL	13.56%	13.52%	87.09%	86.64%

Table A.1: Evaluation results of Task A under the closed track

ID	Team	Task C (with variant characters included)			
		CER	NED	F1	Comprehensive
1	TJU	4.33%	4.33%	95.80%	95.71%
2	SYSU	6.73%	6.72%	93.39%	93.31%
3	WHU-SAI	7.51%	7.50%	92.67%	92.55%
4	NEFU	7.93%	7.91%	92.22%	92.12%
5	SL-ISTIS	7.93%	7.93%	92.16%	92.10%
6	RUC-MIDU	8.23%	8.22%	91.97%	91.83%
7	BNU	9.05%	9.03%	91.33%	91.33%
8	QLNU	8.79%	8.73%	91.40%	91.28%
9	WHU	9.63%	9.61%	90.51%	90.42%
10	NJU	10.10%	10.02%	90.18%	90.00%
11	FDU	13.22%	13.20%	87.23%	86.92%
12	ENC-PSL	21.25%	21.25%	80.12%	79.16%
ID	Team	Task C (with variant characters excluded)			
		CER	NED	F1	Comprehensive
1	TJU	5.67%	5.66%	94.46%	94.37%
2	SYSU	7.84%	7.83%	92.28%	92.20%
3	WHU-SAI	9.02%	9.01%	91.16%	91.04%
4	NEFU	9.17%	9.15%	90.98%	90.88%
5	RUC-MIDU	9.52%	9.51%	90.68%	90.54%
6	SL-ISTIS	9.75%	9.75%	90.34%	90.28%
7	QLNU	10.16%	10.10%	90.03%	89.91%
8	BNU	10.50%	10.49%	89.86%	89.61%
9	WHU	11.02%	11.00%	89.12%	89.03%
10	NJU	11.56%	11.47%	88.73%	88.54%
11	FDU	14.38%	14.36%	86.09%	85.76%
12	ENC-PSL	21.45%	21.45%	79.92%	78.96%

Table A.2: Evaluation results of Task C under the closed track

ID	Team	Task B			
		mAP@[.5:.95]	Micro F1	Macro F1	Avg Match IoU
1	WHU-SAI	59.41%	83.42%	86.86%	76.38%
2	NEFU	51.18%	75.53%	78.79%	77.40%
3	TJU	46.57%	57.42%	71.34%	74.30%
4	ENC-PSL	40.51%	70.83%	67.75%	74.03%
5	FDU	38.89%	64.01%	63.18%	81.19%
6	RUC-MIDU	29.33%	24.14%	37.52%	71.40%
7	WHU	28.02%	29.12%	36.09%	71.04%
8	NJU	27.49%	26.72%	41.62%	67.40%
9	SYSU	26.93%	26.50%	36.76%	68.34%
10	SL-ISTIS	26.19%	19.46%	33.14%	72.25%
11	QLNU	18.58%	4.29%	20.29%	63.90%
12	BNU	14.02%	7.14%	11.73%	70.48%

Table A.3. Evaluation results of Task B under the closed track

A.2 Baselines

To provide a unified performance reference for the participating systems of this competition, current mainstream multimodal large models were selected to construct the baseline systems, namely the general multimodal large model Qwen2.5_VL_7B_Instruction, and Xunzi_Qwen2_VL_7B_Instruction optimized for the ancient document domain. Two processing paradigms were set for the baseline systems: the zero-shot inference paradigm without fine-tuning, and the supervised learning paradigm with instruction fine-tuning based on the training set of this competition, which fully covers the mainstream technical paths that participating teams may adopt. The performance results of each baseline system under all tasks and evaluation rules are shown in Table A.4.

The baseline test results show that the Qwen2.5_VL_7B_Instruction model with instruction fine-tuning on the training set achieved the optimal baseline performance in all tasks. Under the variant-included rule for Task A, the comprehensive score of the fine-tuned model reached 0.9397, a 3.60% improvement compared with the non-fine-tuned state, and the CER decreased from 0.1014 to 0.0618, with a significant performance improvement. Under the variant-included rule for Task C, the comprehensive score of the fine-tuned model increased from 0.8812 to 0.9086, also showing a significant performance gain.

Task A (with variant characters included)					
Model Name	**Fine-tuning Stage**	**CER**	**NED**	**F1**	**Comprehensive**
Qwen2.5_VL_7B_Instruction	No Fine-tuning	10.14%	9.47%	91.10%	90.37%
	Instruction Fine-tuning on Training Set	6.18%	6.13%	94.30%	93.97%
Xunzi_Qwen2_VL_7B_Instruction	No Fine-tuning	17.86%	17.40%	84.09%	82.82%
	Instruction Fine-tuning on Training Set	12.14%	11.83%	89.93%	88.54%
Task A (with variant characters excluded)					
Model Name	**Fine-tuning Stage**	**CER**	**NED**	**F1**	**Comprehensive**
Qwen2.5_VL_7B_Instruction	No Fine-tuning	11.21%	10.54%	90.07%	89.31%
	Instruction Fine-tuning on Training Set	6.85%	6.79%	93.64%	93.31%
Xunzi_Qwen2_VL_7B_Instruction	No Fine-tuning	18.51%	18.02%	83.45%	82.18%
	Instruction Fine-tuning on Training Set	12.64%	12.32%	89.45%	88.05%
Task B					
Model Name	**Fine-tuning Stage**	**mAP@[.5:.95]**	**Micro F1**	**Macro F1**	**Avg Match IoU**
Qwen2.5_VL_7B_Instruction	No Fine-tuning	0.00%	0.00%	0.00%	0.00%
	Instruction Fine-tuning on Training Set	20.06%	5.13%	15.30%	66.00%
Xunzi_Qwen2_VL_7B_Instruction	No Fine-tuning	2.36%	0.03%	1.14%	59.43%
	Instruction Fine-tuning on Training Set	19.17%	4.03%	11.30%	66.54%

Task C (with variant characters included)					
Model Name	**Fine-tuning Stage**	**CER**	**NED**	**F1**	**Comprehensive**
Qwen2.5_VL_7B_Instruction	No Fine-tuning	12.07%	11.93%	88.49%	88.12%
	Instruction Fine-tuning on Training Set	9.20%	9.19%	90.99%	90.86%
Xunzi_Qwen2_VL_7B_Instruction	No Fine-tuning	14.97%	14.92%	85.38%	85.14%
	Instruction Fine-tuning on Training Set	13.83%	13.76%	86.73%	86.35%
Task C (with variant characters excluded)					
Model Name	**Fine-tuning Stage**	**CER**	**NED**	**F1**	**Comprehensive**
Qwen2.5_VL_7B_Instruction	No Fine-tuning	13.38%	13.24%	87.18%	86.81%
	Instruction Fine-tuning on Training Set	10.66%	10.65%	89.53%	89.40%
Xunzi_Qwen2_VL_7B_Instruction	No Fine-tuning	16.09%	16.04%	84.25%	84.02%
	Instruction Fine-tuning on Training Set	15.20%	15.12%	85.34%	84.98%

Table A.4: Baseline evaluation results

Task Name	**Evaluation Rule**	**Modality**	**CER (%)**	**NED (%)**	**F1-score (%)**	**Comprehensive Score (%)**
Task A	With variant characters included	Closed Modality	2.71(-3.47)	2.70(-3.43)	97.54(+3.24)	97.36(+3.39)
		Open Modality	3.03(-3.15)	3.01(-3.12)	97.15(+2.85)	97.03(+3.06)
	With variant characters excluded	Closed Modality	2.75(-4.10)	2.74(-4.05)	97.50(+3.86)	97.32(+4.01)
		Open Modality	3.03(-3.82)	3.01(-3.78)	97.15(+3.51)	97.03(+3.72)
Task C	With variant characters included	Closed Modality	4.33(-4.87)	4.33(-4.86)	95.80(+4.81)	95.71(+4.85)
		Open Modality	5.52(-3.68)	5.49(-3.70)	94.71(+3.72)	94.56(+3.70)
	With variant characters excluded	Closed Modality	5.67(-4.99)	5.66(-4.99)	94.46(+4.93)	94.37(+4.97)
		Open Modality	7.03(-3.63)	7.00(-3.65)	93.20(+3.67)	93.05(+3.65)
Task Name	**Evaluation Rule**	**Modality**	**mAP@[.5:.95]**	**Micro F1**	**Macro F1**	**Avg Match IoU**
Task B	-	Closed Modality	59.41(+39.35)	83.42(+78.29)	86.86(+71.56)	76.38(+10.38)
		Open Modality	58.18(+38.12)	82.78(+77.65)	86.45(+71.15)	75.72(+9.72)

Table A.5: Performance improvement of the optimal participating results wrt the baseline model (%)

Although the Xunzi_Qwen2_VL_7B_Instruction model has been pre-trained and optimized for the ancient document domain, its performance is lower than that of the general multimodal large model Qwen2.5_VL_7B_Instruction in both zero-shot and fine-tuning scenarios of this task. The core reason is that the pre-training focus of this model is concentrated on the semantic understanding of ancient document texts, and it has obvious shortcomings in the visual feature learning of ancient document images and character recognition ability.

To quantify the performance improvement of the participating systems relative to the baseline model, the optimal participating results under each task and each modality were selected to compare with the performance of the core baseline model. Based on the average value of the baseline data, the maximum improvement range and specific values of each core indicator were calculated, and the results are shown in T able A.5.

The performance improvement results show that the optimal performance of all participating systems in this competition is significantly better than that of the official baseline model. For the character recognition tasks, the overall performance improvement under the closed modality is higher than that under the open modality, with the improvement range of comprehensive score from 3.06% to 4.97%, and the maximum reduction range of CER from 3.15% to 4.99%. For the layout analysis task, the performance improvement of the participating systems is more significant, with the improvement range of mAP@[.5:.95] from 38.12% to 39.35%, and the improvement range of Macro F1-score from 71.15% to 71.56%. This reflects that the model optimization and adaptation strategies adopted by the participating teams, targeting the domain characteristics of ancient document visual processing, are extremely effective and can significantly break through the performance bottleneck of general multimodal large models in ancient document layout recognition tasks.

A.3 Analysis of Variant Character Recognition and Character Distribution Characteristics

Rank	Char	Frequency	%
1	無	2,694	1.66%
2	之	1,829	1.13%
3	不	1,786	1.10%
4	一	1,513	0.93%
5	為	1,473	0.91%
6	以	1,411	0.87%
7	空	1,305	0.80%
8	有	1,188	0.73%
9	二	1,179	0.73%
10	所	1,154	0.71%

Table A.6: Statistics of top 10 high-frequency characters across datasets

Table A.6 presents the frequency statistics of the top 10 high-frequency characters across all datasets in this evaluation. Collectively, these 10 characters amount to 15,532 tokens, accounting for 9.57% of the total 162,332 character tokens in the full corpus. This distribution explicitly reveals the pronounced long-tail feature and severe class imbalance inherent in ancient Chinese texts. While models achieve stable and high recognition accuracy on these prevalent characters with extensive pre-training coverage, low-frequency rare characters and non-standard variant forms, which are particularly concentrated in handwritten datasets, remain the core bottleneck restricting the overall performance of ancient Chinese OCR systems.

Appendix B: Character Distribution Statistics

The following tables present the full statistical distribution of Chinese characters across Unicode blocks for all dataset splits, referenced from Section 6.2.

Unicode Block	Train_A	Test_A	Train_C	Test_C
CJK Basic	81,626 (97.80%)	4,540 (99.65%)	67,302 (97.85%)	4,716 (99.75%)
Ext-B	1,664 (1.99%)	13 (0.29%)	695 (1.01%)	0 (0.00%)
Ext-A	165 (0.20%)	1 (0.02%)	469 (0.68%)	12 (0.25%)
Ext-F	1 (0.00%)	0 (0.00%)	213 (0.31%)	0 (0.00%)
Ext-E	2 (0.00%)	0 (0.00%)	36 (0.05%)	0 (0.00%)
Ext-D	0 (0.00%)	0 (0.00%)	35 (0.05%)	0 (0.00%)
Ext-C	2 (0.00%)	0 (0.00%)	33 (0.05%)	0 (0.00%)
Compati-bility Supple-ment	0 (0.00%)	2 (0.04%)	0 (0.00%)	0 (0.00%)
Sum	83,460 (100.00%)	4,556 (100.00%)	68,783 (100.00%)	4,728 (100.00%)

Table B.1: Statistical Distribution by Total Character Frequency

Unicode Block	Train_A	Test_A	Train_C	Test_C
CJK Basic	3,162 (98.20%)	1,406 (99.58%)	4,523 (96.01%)	680 (99.85%)
Ext-B	33 (1.02%)	3 (0.21%)	95 (2.02%)	0 (0.00%)
Ext-A	21 (0.65%)	1 (0.07%)	51 (1.08%)	1 (0.15%)
Ext-F	1 (0.03%)	0 (0.00%)	25 (0.53%)	0 (0.00%)
Ext-E	2 (0.06%)	0 (0.00%)	7 (0.15%)	0 (0.00%)
Ext-D	0 (0.00%)	0 (0.00%)	6 (0.13%)	0 (0.00%)
Ext-C	1 (0.03%)	0 (0.00%)	4 (0.08%)	0 (0.00%)
Compa-tibility Supple-ment	0 (0.00%)	2 (0.14%)	0 (0.00%)	0 (0.00%)
Total Sum	3,220 (100.00%)	1,412 (100.00%)	4,711 (100.00%)	681 (100.00%)

Table B.2: Statistical Distribution by Unique Character Vocabulary

Appendix C: Auxiliary Evaluation Metrics

This appendix details the auxiliary metrics used to provide a more granular analysis of the models' performance across Tasks A, B, and C, as well as the comprehensive score used for final rankings.

C.1 Auxiliary Metrics for Text Recognition (Tasks A and C)

(1)Macro and Micro CER:

To prevent a single metric from introducing evaluation bias, we further split CER into Macro CER and Micro CER:

$$\text{Macro CER}=\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\text{CER}_{\mathrm{i}} \qquad (C\text{-}1)$$

$$\text{Micro CER}=\frac{\sum_{\mathrm{i}=1}^{\mathrm{N}}\text{Edit\_Distance}(\text{ref}_{\mathrm{i}},\text{hyp}_{\mathrm{i}})}{\sum_{\mathrm{i}=1}^{\mathrm{N}}|\text{ref}_{\mathrm{i}}|} \qquad (C\text{-}2)$$

(2) Precision, Recall, and F1-score

Based on the number of correctly matched characters, we calculate three auxiliary metrics that intuitively reflect the model's comprehensive ability in positive sample recognition:

$$\text{Precision}=\frac{\mathrm{N}_{\text{correct}}}{|\text{hyp}|} \qquad (C\text{-}3)$$

$$\text{Recall}=\frac{\mathrm{N}_{\text{correct}}}{|\text{ref}|} \qquad (C\text{-}4)$$

$$\text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \qquad (C\text{-}5)$$

where $N_{correct}$ is the number of correctly matched characters between the predicted text and the reference text.

(3) Normalized Edit Distance (NED)

Considering that conventional edit distance can easily fail when there is a drastic length deviation between the predicted text and the ground truth label, we supplemented the Normalized Edit Distance (NED). NED ensures evaluation robustness under extreme length differences by using the maximum length of the two as the denominator:

$$\text{NED}=\frac{\text{Edit\_Distance}(\text{ref},\text{hyp})}{\max(|\text{ref}|,|\text{hyp}|)} \qquad (C\text{-}6)$$

(4) Comprehensive Score

To provide an intuitive basis for the final ranking, we designed a weighted comprehensive score integrating the above dimensions. This score blends the character error rate, exact match rate, and normalized edit distance in different proportions, and its value is strictly mapped to the range [0,1]. The calculation formula is:

$$\text{Score}_{\text{comprehensive}}=$$

$$(1\text{-CER})\times0.5+\text{F1}\times0.3+(1\text{-NED})\times0.2 \qquad (C\text{-}7)$$

A higher score indicates better comprehensive model performance. This metric is used to determine the final ranking of participating teams on datasets A and C.

C.2 Auxiliary Metrics for Object Detection (Task B)

(1)F1 Score

Both Macro-F1 and Micro-F1 calculation methods are provided. For the label set $\mathbb{L}$, they are defined as:

$$\text{Micro-F1}=\frac{2\times\sum_{\mathrm{l}\in\mathbb{L}}\text{TP}_{\mathrm{l}}}{2\times\sum_{\mathrm{l}\in\mathbb{L}}\text{TP}_{\mathrm{l}}+\sum_{\mathrm{l}\in\mathbb{L}}\text{FP}_{\mathrm{l}}+\sum_{\mathrm{l}\in\mathbb{L}}\text{FN}_{\mathrm{l}}} \qquad (C\text{-}8)$$

$$\text{Macro-F1}=\frac{1}{|\mathbb{L}|}\sum_{\mathrm{l}\in\mathbb{L}}\frac{2\times\text{TP}_{\mathrm{l}}}{2\times\text{TP}_{\mathrm{l}}+\text{FP}_{\mathrm{l}}+\text{FN}_{\mathrm{l}}} \qquad (C\text{-}9)$$

where TP_l, FP_l, and FN_l are the number of true positives, false positives, and false negatives for label l, respectively.

(2) Mean IoU of Matched Boxes

When determining whether a predicted box and a ground truth annotation box are successfully matched, we use Intersection over Union (IoU) as the core basis. The algorithm employs a greedy strategy to group labels within the same image, preferentially matches the box pair with the maximum IoU value, and considers IoU 0.5 as a valid detection. Its calculation formula is:

$$\overline{\text{IoU}}=\frac{1}{\mathrm{N}_{\text{match}}}\sum_{(\mathrm{B}_{\mathrm{p}},\mathrm{B}_{\mathrm{g}})\in\mathbb{M}}\text{IoU}\,(\mathrm{B}_{\mathrm{p}},\mathrm{B}_{\mathrm{g}}) \qquad (C\text{-}10)$$

where $\mathbb{M}$ is the set of successfully matched box pairs, and $N_{match}=|\mathbb{M}|$.

Appendix D: Error Analysis and Discussion

D.1 Failures in Complex Layouts

The typesetting and layout of ancient Chinese documents differ significantly from modern printing. Among these differences, complex elements such as interlinear small-character annotations (Shuanghang Xiaozhi), raised items (Taixiang), seals, and book margins (Shukou) pose severe challenges to the model's layout analysis (Task B). Evaluation results indicate that even the best-performing system in the closed track achieved an average precision (mAP@[.5:.95]) of only 0.5941 in layout analysis, which is far below the performance level of character recognition tasks.

Among the numerous factors causing layout analysis failures, the confusion between double-line annotations and the main text is the most prominent. From the perspective of visual

features, interlinear small-character annotations in ancient books are tightly embedded between single columns of large-character main text, forming a unique "interlocking text-and-image nesting" phenomenon. Most existing object detection models or layout analysis paradigms are typically pre-trained on modern documents (e.g., left-to-right, horizontal layout, distinct line spacing). When processing ancient book images, models are highly prone to structural misjudgments: because the physical distance between the large main text characters and the small annotation characters is too close, the model easily groups the entire column (containing both main text and annotations) into a single Text Block, resulting in the failure of fine-grained layout element segmentation. Furthermore, although some models perform well on the Avg Match IoU metric for layout region localization, they fail to accurately separate “text” (main text) from "annotations" according to typographic logic, leading to a drastic decline in comprehensive classification metrics (such as Macro F1). Future structural parsing should integrate mature document layout frameworks, such as LayoutParser (Shen et al., 2021), to mitigate typographic logic errors.

D.2 Variant and Rare Characters

One of the core difficulties in optical character recognition (OCR) lies in the accurate recognition of long-tail rare characters and variant characters (Wang et al., 2019). Although variant characters exert a measurable influence on system performance, their relatively low occurrence frequency means that the overall impact on evaluation scores remains limited. To quantitatively examine this effect, this study conducted a statistical mapping of Chinese character Unicode blocks for Dataset A (block-engraved editions) and Dataset C (manuscripts) from two dimensions: Total Character Frequency and Unique Character Vocabulary (deduplicated non-repeating characters). The detailed statistical results are provided in Appendix B (Tables B1 and B2).

The statistical results reveal the extreme long-tailed distribution characteristics of ancient Chinese texts. In the dimension of character frequency (Table B1), the CJK Basic block (U+4E00–U+9FFF) maintains absolute dominance across all subsets (97.80% to 99.75%). This explains why even baseline models without domain-specific fine-tuning can often achieve a basic recognition rate of around 90%.

Regarding the recognition performance of characters in Unicode extension blocks, our analysis reveals a pronounced performance gap compared to CJK Basic characters. Characters from the CJK Basic block (U+4E00-U+9FFF) are recognized with high accuracy owing to their high frequency and broad coverage in model pre-training corpora. In contrast, characters belonging to extension blocks (Ext-A through Ext-F) exhibit substantially higher error rates. For instance, in the manuscript dataset (Dataset C), extension-block characters account for approximately 4% of the unique character vocabulary; yet their CER contribution is disproportionately large relative to their frequency. This disparity is particularly pronounced for Ext-B characters (U+20000-U+2A6DF), which represent obscure historical glyphs rarely encountered in modern digitized corpora. These findings confirm that improving the coverage of extension-block characters in model training data and vocabulary design is a critical direction for advancing ancient Chinese OCR systems.

However, in the dimension of unique character vocabulary (Table B2), different text carriers present significant variances and recognition challenges. Compared to the well-standardized block-engraved editions (Train_A), the unique character vocabulary (vocabulary diversity) of the manuscript texts (Train_C) expands sharply to 4,711. More importantly, in Train_C, variant and rare characters located in the extension blocks outside the CJK Basic set (Ext-A to Ext-F) contribute to nearly 4% of the character vocabulary diversity, with a distribution breadth far exceeding that of Train_A. For instance, Ext-B and Ext-F account for 2.02% and 0.53% of the unique characters in the manuscripts, respectively.

This data structural contrast—“extremely rich in unique characters but extremely low in occurrence frequency”—constitutes the core bottleneck for current large multimodal models(LMMs) in ancient text OCR tasks. Because the visual encoders of open-source LMMs (such as Qwen2.5-VL) rarely cover low-frequency rare characters in the Ext-B to Ext-F blocks during the pre-training phase, models inevitably encounter severe representation dimensionality reduction or Out-of-Vocabulary (OOV) truncation when processing such long-tail characters. This causes the Character Error Rate (CER) for the long-tail character segment to remain persistently high.

Beyond the Unicode distribution analysis, we further conducted a fine-grained evaluation of how each participating team’s system actually handled the specific variant character instances present in the test sets. Using the predefined variant character mapping table (19 mapping pairs in total), we identified all positions in the ground-truth annotations where a variant character occurred and examined whether each team's prediction was correct at those positions. Two accuracy metrics are reported: Strict Accuracy, which counts a prediction as correct only when the predicted character is identical to the ground-truth variant character; and Loose Accuracy, which additionally accepts the

equivalent standard-form counterpart as correct. Table D.1 summarizes the results across all thirteen participating teams.

Team	Task A Strict Acc.	Task A Loose Acc.	Task C Strict Acc.	Task C Loose Acc.
BNU	85.3%	94.1%	56.8%	79.3%
ENC-PSL	47.1%	77.9%	20.6%	47.1%
EPHE-PSL	94.1%	95.6%	49.7%	72.9%
FDU	83.8%	88.2%	61.9%	84.5%
NEFU	76.5%	89.7%	36.8%	78.7%
NJU	60.3%	92.1%	48.4%	48.4%
QLNU	95.6%	95.6%	46.5%	83.2%
RUC-MIDU	95.6%	95.6%	40.6%	71.6%
SL-ISTIS	94.1%	95.6%	63.2%	92.9%
SYSU	70.6%	80.9%	30.3%	65.8%
TJU	91.2%	94.1%	40.6%	68.4%
WHU	95.2%	95.2%	49.7%	79.3%
WHU-SAI	94.1%	94.1%	47.7%	77.4%

Table D.1: Variant character recognition accuracy

Several patterns emerge from Table 12. First, Task A variant accuracy is substantially higher than Task C across all teams, with most teams achieving loose accuracy above 88% on Task A, while Task C loose accuracy ranges widely from 47.1% (ENC-PSL) to 92.9% (SL-ISTIS). This performance gap is consistent with the broader finding that handwritten manuscripts pose greater recognition difficulties than block-engraved texts, and specifically reflects that the variant character forms encountered in handwriting are more visually irregular and diverse.

Second, the gap between strict and loose accuracy reveals how systems handle graphically equivalent forms. Teams such as NEFU (Task C: 36.8% strict vs. 78.7% loose, a gap of 41.9 percentage points) and QLNU (Task C: 46.5% vs. 83.2%, a gap of 36.7 points) show that their models frequently predict the standard-form counterpart of a variant character rather than the variant itself. This reflects the strong prior toward modern standardized character forms instilled during pre-training, which causes models to unconsciously "normalize" the visual input even when the ground truth requires the historical variant. The character 稅/税 is a representative example: the ground-truth character in Dataset C is 稅 (the variant form), yet the majority of teams predominantly predicted 税, the modern standard form, resulting in high loose accuracy but low strict accuracy.

Third, the character 関 proved to be the most challenging variant across all teams in Task C. It was frequently confused with visually similar but unrelated characters such as 閩, 聞, and 閱, leading to low loose accuracy for many systems (e.g., ENC-PSL: 1/47, NJU: 11/47). Unlike the 税/稅 case - where the error pattern is systematic and semantically grounded - errors on 関 appear more perceptually driven, suggesting that models fail to reliably distinguish this character's distinctive graphemic features from surrounding similar glyphs.

D.3 Systematic Biases

The adoption of LMMs as the dominant technical backbone in this evaluation—employed by virtually all participating teams across both modalitie—brings into focus two systematic failure modes that are distinct in mechanism yet share a common root: the conflict between the model's intrinsic language prior, shaped by pre-training on modern standardized text, and the visual fidelity demanded by archival OCR. These two modes are hallucination and orthographic over-normalization, and the results of this evaluation provide concrete evidence for both.

Hallucination: perceptual failure driven by language prior. Hallucination in the OCR context refers to cases where the model generates a character that bears no faithful correspondence to the visual content of the image, where the language decoder, failing to obtain a reliable signal from the visual encoder, defaults to a contextually plausible character rather than the one actually present. The clearest evidence in our results comes from the character 関 in Dataset C. Across all thirteen participating teams, 関 was the most poorly recognized variant character, with loose accuracy as low as 1/47 (ENC-PSL) and 11/47 (NJU). Crucially, the errors were not random: the character was systematically confused with visually similar but semantically unrelated characters - 閩, 聞, 閱 - all of which belong to the high-frequency CJK Basic block and appear extensively in classical Chinese pre-training corpora. The pattern is precisely what hallucination predicts: when the visual encoder produces an uncertain representation of an unfamiliar or low-frequency glyph, the language model component intervenes and "completes" the output toward a character it has seen many times in similar syntactic contexts, overriding the visual evidence. This mechanism is further amplified for characters in Unicode extension blocks (Ext-B through Ext-F), which are rarely encountered in modern digitized corpora and for which the visual decoder has no stable internal representation; in such cases, the language prior becomes the only signal effectively, producing outputs that may be linguistically coherent but graphically unfaithful.

Orthographic over-normalization: generation-stage bias toward standard forms. The second failure mode is subtler and, from the perspective of documentary fidelity, equally damaging. Over-

normalization does not involve a failure of visual perception: the model correctly identifies the semantic identity of the character in the image, but in the generation stage substitutes the modern standard form for the historical variant actually depicted. The gap between strict and loose accuracy observed across teams in Dataset C provides systematic evidence for this phenomenon. Teams such as NEFU (strict: 36.8%, loose: 78.7%, gap: 41.9 pp) and QLNU (strict: 46.5%, loose: 83.2%, gap: 36.7 pp) show that a large proportion of their "errors" under strict evaluation are in fact correct recognitions of the character's semantic identity, delivered in the wrong orthographic form. The character pair 稅/稅 is paradigmatic: the ground truth in Dataset C is 稅 (the historical variant), yet the overwhelming majority of teams predicted 稅 (the modern standard form), because this form dominates the pre-training corpus and is thus assigned a higher generation probability regardless of what is visually present. Unlike hallucination, over-normalization preserves semantic meaning while erasing graphemic authenticity—a trade-off that is acceptable in semantic NLP tasks but constitutes a fundamental failure in archival transcription, where the exact character form carries philological and paleographic significance.

Shared root, different remedies. Both failure modes originate from the same structural asymmetry: pre-training corpora for large models consist overwhelmingly of collated, standardized modern text, leaving rare glyphs, extension-block characters, and historical variant forms severely underrepresented. However, they require different interventions. Hallucination, arising from visual encoder uncertainty, calls for stronger visual grounding—improving the coverage of rare and archaic glyphs in pre-training, and introducing visual confidence mechanisms that prevent the language decoder from overriding low-confidence visual representations. Over-normalization, arising from generation-stage bias, calls for orthographic calibration—through fine-tuning on corpora that explicitly include historical variant forms, alignment training objectives (such as DPO) that penalize substitution of variants with their standard counterparts, or constrained decoding strategies that restrict the output vocabulary to forms attested in the source image's character set. The results of this evaluation suggest that fine-tuning on in-domain data partially addresses both problems—top closed-track systems exhibit neither failure mode at a significant rat—but the underlying biases are not eliminated, and will resurface whenever these models are applied to document collections outside their fine-tuning distribution.

D.4 A New Framework for Ancient Chinese OCR & Layout Analysis

The task of recognizing text from ancient Chinese document images can be approached through two fundamentally distinct architectural paradigms. The first is an end-to-end paradigm, in which a single multimodal large model processes the raw page image and directly outputs the transcribed text, performing implicit layout understanding and character recognition in a unified forward pass. The second is a pipeline paradigm, in which layout analysis and OCR are decoupled into sequential stages: a layout analysis model first identifies and segments text regions (columns, annotations, main body), after which a dedicated OCR component transcribes each region independently.

Both paradigms were represented among the participating systems in this evaluation, and their respective strengths and limitations are revealed by the results.

The end-to-end approach, exemplified by systems built directly on Qwen2.5-VL, offers substantial practical advantages: a simpler engineering stack, no dependency on intermediate layout annotations, and the ability to leverage the model's broad world knowledge for contextual disambiguation. The best-performing closed-modality system overall — TJU, which achieved a Task A comprehensive score of 0.9736 and a Task C score of 0.9571 — demonstrates that a well-optimized end-to-end architecture can reach state-of-the-art performance. However, the end-to-end approach is also more susceptible to the over-generation failure mode described in Section 6.3: without explicit layout constraints, the model must infer reading order, column boundaries, and the distinction between main text and interlinear annotations from visual context alone, which can lead to systematic errors in structurally complex pages.

The pipeline approach, exemplified by TJU's integration of HistLayout-DETR for layout pre-segmentation prior to OCR, addresses this weakness by making layout structure explicit. When text regions are correctly isolated, the downstream OCR component operates on substantially simpler inputs - individual columns or lines rather than full page images-reducing ambiguity and limiting the effective receptive field over which over-generation can occur. The results of WHU-SAI in Task B (mAP of 0.5941, the highest in the evaluation) further suggest that dedicated layout models, trained specifically on the structural peculiarities of ancient Chinese books, can achieve precision that end-to-end models currently cannot match. This structural advantage is particularly valuable for documents containing double-line annotations (Shuanghang Xiaozhi), raised items (Taixiang), and other

complex typographic elements whose correct segmentation is a prerequisite for accurate transcription.

The trade-off, however, is that the pipeline approach introduces error propagation: layout segmentation error—misclassified regions, merged columns, or missed annotation blocks—directly degrade the quality of downstream OCR in ways that are difficult to recover from. Moreover, pipeline systems require labeled layout data for training the segmentation stage, which is costly to produce for historical documents. The baseline results (Table 8) illustrate this brittleness: the zero-shot Qwen2.5-VL baseline achieves an mAP of 0.00 on Task B without fine-tuning, indicating that general-purpose models have virtually no out-of-the-box layout understanding for ancient Chinese books.

Looking forward, we propose that the most promising direction is not a choice between these two paradigms, but a staged hybrid architecture that combines their respective strengths. Concretely, such a system would consist of three components: (1) a lightweight layout analysis module trained on domain-specific annotations to segment the page into typed regions; (2) a reading-order resolution module that reconstructs the logical reading sequence—right-to-left, top-to-bottom, with correct interleaving of main text and annotation—from the detected regions; (3) a fine-tuned multimodal OCR model that transcribes each region independently, operating on clean, pre-segmented inputs rather than full page images. This architecture preserves the recognition power of LMMs while anchoring their outputs to explicit structural constraints, thereby suppressing over-generation and improving robustness on the complex typographic elements that currently constitute the primary bottleneck for ancient Chinese document OCR systems.