A Dual-Modality Framework for Ancient Document Layout Analysis and Text Recognition

Qi Fan*, **Jieming Hu***, **Chen Ye**†
Department of Computer Science and Technology, Tongji University, Shanghai, China
*These authors contributed equally to this work.
†Corresponding author.
{2350262, 2351584, yechen}@tongji.edu.cn

Abstract

The digital preservation of ancient Chinese literature requires robust capabilities spanning layout analysis and text recognition. This paper presents a comprehensive framework addressing two fundamental challenges: (1) **Layout Element Analysis (Task B)** for detecting page elements (text, image, book_edge, seal) amidst degradation, nested structures, and extreme class imbalance; and (2) **Text Recognition (Tasks A & C)** for end-to-end transcription of printed and handwritten classical documents. For layout analysis, we propose a dual-modality solution. The Closed Modality formulates this as a sequence-to-sequence problem using Vision-Language Models (VLMs), introducing spatial discretization tokenization and a Frequency-Aware Sequential Curriculum Learning framework with dynamic memory replay. The Open Modality presents HistLayout-DETR, a set prediction architecture integrating an Augmented Morphological Encoder and a Polygon Boundary Refinement head. For text recognition, we formulate OCR as a domain-constrained visual language generation task using Qwen2.5-VL with LoRA fine-tuning. We employ structured prompts encoding reading order and Traditional Chinese character preservation across domains. Extensive experiments on the EvaHan 2026 dataset validate our framework's superiority. In layout analysis, our curriculum-guided paradigm achieves a Macro F1 of 0.7992 and mAP@[.5:.95] of 0.5438. In text recognition, we achieve CERs of 0.0271 on printed and 0.0433 on handwritten texts.

Keywords: Ancient Document Layout Analysis, Text Recognition, Vision-Language Models, HistLayout-DETR, Curriculum Learning, Polygon Regression

1. Introduction

The digital preservation of ancient Chinese literature relies heavily on precise Layout Element Analysis and Text Recognition. The EvaHan 2026 benchmark introduces rigorous tasks for layout parsing (Task B) and OCR for printed and handwritten texts (Tasks A & C). However, extracting structural and semantic elements from historical scans remains profoundly challenging. Key obstacles include a severe long-tailed distribution of page elements (where text vastly outnumbers seals or book edges), slanted and nested non-axis-aligned structures, the absence of chromatic cues in degraded black-and-white (B&W) scans, vertical right-to-left reading orders, and the intricate preservation of rare Traditional Chinese characters across highly variable handwritten calligraphy.

To comprehensively address these challenges, we propose a robust, unified framework tailored for historical document understanding.

For **Layout Analysis (Task B)**, we present a dual-modality framework. In the **Closed Modality**, we formulate the task as an autoregressive sequence generation problem utilizing the Qwen2.5-VL backbone. We introduce a spatial discretization tokenization protocol to resolve coordinate scaling issues, alongside a novel Frequency-Aware Sequential Curriculum Learning framework. By integrating a dynamic curriculum replay loss with temporal decay and frequency-inverse penalties, our approach successfully mitigates dominant-class bias and catastrophic forgetting. For the **Open Modality**, we propose **HistLayout-DETR**, an end-to-end set prediction architecture. Meticulously engineered with an Augmented Morphological Encoder (featuring Adaptive Grain-Size Attention and Morphology-Aware Texture Enhancement), a Hierarchical Text-in-Image Decoder, and a Polygon Boundary Refinement (PBR) head, it excels at extracting multi-scale structural signatures and regressing precise, non-axis-aligned quadrilateral boundaries.

For **Text Recognition (Tasks A & C)**, we formulate classical Chinese OCR as a domain-constrained visual language generation task using large Vision-Language Models (VLMs). We employ a structured prompting framework that explicitly encodes reading order, script preservation, and uncertainty handling. Joint training across printed and handwritten domains enables the model to learn shared semantic priors while adapting to severe visual variability.

2. Related Work

Deep learning architectures, ranging from Convolutional Neural Networks (Xu et al., 2018; Redmon et al., 2016) to Transformers (Zhu et al., 2021; Banerjee et al., 2023), have significantly advanced historical document layout analysis and Chinese OCR (Shi et al., 2018; Baek et al., 2021). However,

Proceedings of LT4HALA 2026 @ LREC 2026, pages 277–287.

traditional layout detectors typically regress axis-aligned bounding boxes, which are inadequate for the slanted columns and nested structures (e.g., text within illustrations) prevalent in ancient scans. Recently, Vision-Language Models (VLMs) (Bai et al., 2023; Kim et al., 2022) have unified visual detection and text recognition into autoregressive generation tasks, inspired by spatial coordinate quantization (Chen et al., 2022). Despite their impressive capabilities (Liu et al., 2023; Li et al., 2023), directly applying VLMs to historical archives introduces severe Out-of-Vocabulary (OOV) errors due to varying digitization scales. Furthermore, ancient datasets exhibit extreme long-tailed distributions. While curriculum learning (Bengio et al., 2009) and continual adaptation (Rebuffi et al., 2017) can address data imbalance, naive sequential fine-tuning inevitably triggers catastrophic forgetting (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017), and standard static memory replay often reinforces dominant-class bias.

3. Layout Analysis (Task B)

We formulate Layout Element Analysis (Task B)—detecting text, image, book_edge, and seal from an ancient document image $I \in \mathbb{R}^{H\times W\times 3}$ into bounding boxes B and labels C—via a dual-modality approach. For the Open Modality, we propose a specialized DETR-based architecture. For the Closed Modality, we cast the task as an autoregressive sequence-to-sequence generation problem using Vision-Language Models (VLMs). To efficiently fine-tune the VLM baseline, we apply Low-Rank Adaptation (LoRA) by injecting trainable rank decomposition matrices (rank=32, alpha=64, target_modules=all-linear) into the frozen backbone's transformer layers. The model is trained using bfloat16 precision for 3 epochs with a learning rate of 5e-5 and a 0.1 warmup ratio.

3.1. Paradigm I: Joint Instruction Fine-Tuning with Discretized Coordinate Normalization

In Paradigm I, we aggregate instances of all four layout categories for simultaneous joint fine-tuning. To enable the VLM backend to natively process spatial bounding boxes without architectural modifications, we map continuous, unbounded raw pixel coordinates into a standardized, resolution-agnostic spatial grid of $[0, 1000]$ discrete text tokens (see Figure 1). Directly feeding raw coordinates (e.g., $x = 219, y = 60$) induces severe out-of-vocabulary (OOV) issues and hinders the learning of generalized spatial relationships. By contrast, bounding coordinates to 1000 discrete bins creates a finite, learnable vocabulary of spatial tokens. This normalization effectively decouples geometric layouts from absolute physical dimensions to ensure resolution invariance, while striking an optimal density balance—fine enough to capture the minute details of dense text lines, yet coarse enough to prevent excessive embedding sparsity.

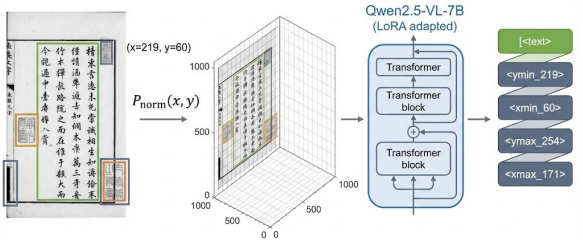

Figure 1: The proposed Spatial Discretization and Tokenization Framework (Paradigm I).

During data preprocessing, every absolute coordinate point $P(x, y)$ in the original image of width W and height H is normalized to $P_{norm}(x_{norm}, y_{norm})$:

$$\begin{aligned} x_{norm} &= \text{round}\left(\frac{x}{W} \times 1000\right) \\ y_{norm} &= \text{round}\left(\frac{y}{H} \times 1000\right) \end{aligned} \tag{1}$$

These normalized values are formatted into text prompts (e.g., [ymin, xmin, ymax, xmax]). During inference, the model autoregressively generates the normalized coordinate tokens. To map the predictions back to the original image space for evaluation, we apply an inverse coordinate restoration:

$$\begin{aligned} \hat{x} &= \text{round}\left(\hat{x}_{norm} \times \frac{W}{1000}\right) \\ \hat{y} &= \text{round}\left(\hat{y}_{norm} \times \frac{H}{1000}\right) \end{aligned} \tag{2}$$

3.2. Paradigm II: Frequency-Aware Sequential Curriculum Learning

Ancient layout elements exhibit a severe long-tailed distribution, causing standard joint training (Paradigm I) to over-optimize for the hyper-abundant text class while neglecting sparse minority elements (e.g., seals). To systematically address this bias, we propose a Frequency-Aware Sequential Curriculum Learning framework (Figure 2). We partition the dataset into class-specific subsets $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_4\}$ ordered by monotonically decreasing instance frequency ($N(c_1) > \cdots > N(c_4)$). Training proceeds sequentially over $K = 4$ stages, enabling the VLM to build a robust global structural foundation before specializing in localized, nuanced elements.

To mitigate catastrophic forgetting during sequential adaptation, we formulate a Dynamic Curriculum Replay Loss. At stage k, the model optimizes over

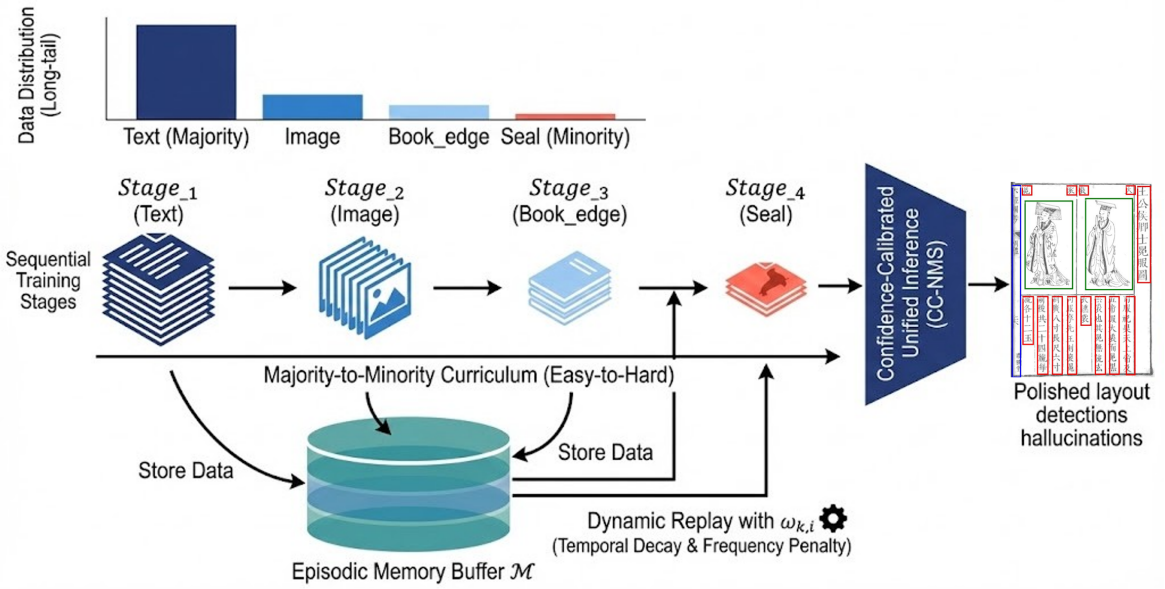

Figure 2: Frequency-Aware Sequential Curriculum Learning with Dynamic Memory Replay. To mitigate long-tail imbalance, training progresses sequentially from majority elements (text) to minority ones (seal).

both the target dataset $\mathcal{D}_k$ and an episodic memory buffer $\mathcal{M}_i$:

$$\mathcal{L}_{\text{total}}^{(k)}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_k}\left[\mathcal{L}_{\text{CE}}(f_\theta(x), y)\right] + \sum_{i=1}^{k-1} \omega_{k,i} \cdot \mathbb{E}_{(x_m,y_m)\sim\mathcal{M}_i}\left[\mathcal{L}_{\text{CE}}(f_\theta(x_m), y_m)\right] \tag{3}$$

The dynamic retention weight $\omega_{k,i}$ for historical class i is defined as:

$$\omega_{k,i} = \lambda \cdot \exp(-\gamma(k-i)) \cdot \left(\frac{N(c_i)}{\max_{j<k} N(c_j)}\right)^{-\rho} \tag{4}$$

This dynamic weight balances plasticity and stability via a base replay coefficient λ, a temporal decay term $\exp(-\gamma(k-i))$ that relaxes the preservation of older concepts to free up parameters, and a frequency-inverse penalty governed by ρ. Crucially, this penalty down-weights historical classes with massive original sample sizes, preventing their replay gradients from swamping the optimization of minority classes.

Finally, because sequential training inherently biases confidence toward recent classes, we enforce confidence-calibrated unified inference. For a generated sequence $\hat{y}$ composed of coordinate tokens $\{v_1, \ldots, v_L\}$, we compute a length-normalized confidence score:

$$S_{\text{conf}}(\hat{y}|x) = \exp\left(\frac{1}{L}\sum_{t=1}^{L} \log P(v_t|v_{<t}, x; \theta^*)\right) \tag{5}$$

A predicted bounding box for class c_k is accepted only if $S_{\text{conf}}(\hat{y}|x) \geq \tau_{c_k}$, effectively filtering out hallucinations while retaining high-recall detection capabilities.

4. Text Recognition (Tasks A & C)

We address text recognition for both printed (Task A/keben) and handwritten (Task C/xieben) ancient Chinese documents by formulating OCR as a domain-constrained visual language generation problem. Given a document image $I \in \mathbb{R}^{H\times W\times 3}$, we utilize the Qwen2.5-VL-7B-Instruct vision-language model to autoregressively generate a text sequence T (Figure 3). To ensure historical validity, the generated sequence must preserve Traditional Chinese characters and adhere to a vertical right-to-left reading order. This constrained generation is modeled as:

$$\hat{T} = \arg\max_T P(T|I, C; \theta) \tag{6}$$

where C denotes explicit domain constraints restricting the output space. Rather than relying on generic instructions, we enforce these constraints via a unified structured prompt that explicitly dictates the reading direction, prohibits Simplified Chinese conversion, and mandates the use of '#' for unrecognizable characters.

To adapt the base model efficiently without catastrophic forgetting, we employ Low-Rank Adaptation (LoRA). Instead of updating the entire network, we freeze the pre-trained weights and inject trainable rank decomposition matrices into the transformer layers:

$$W' = W + \Delta W = W + BA \tag{7}$$

where $W \in \mathbb{R}^{d\times d}$ is the original frozen weight matrix, and $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times d}$ are the trainable matrices. We configure LoRA with rank $r = 32$

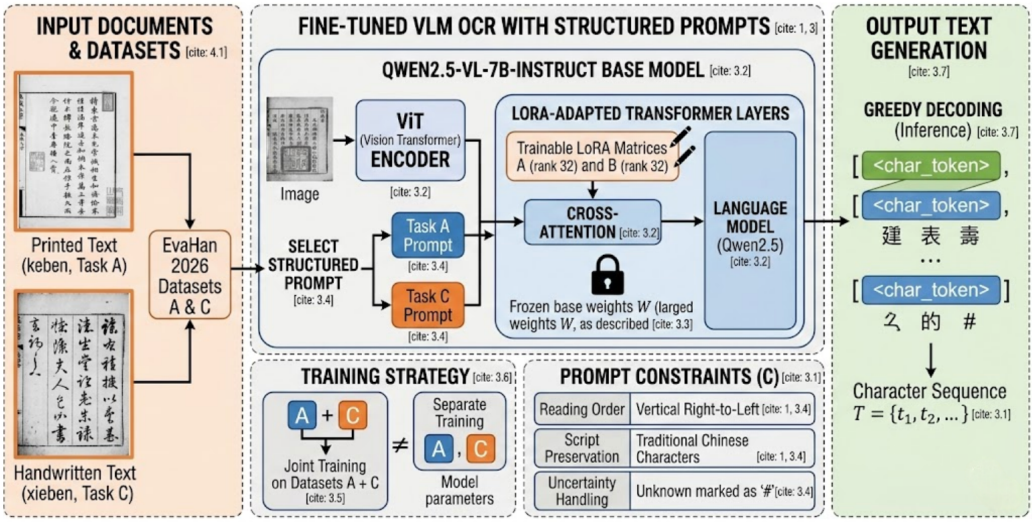

Figure 3: The proposed VLM-based text recognition framework. Historical document images are processed by Qwen2.5-VL with LoRA fine-tuning and structured prompts for domain-constrained OCR.

and $\alpha = 32$, targeting all linear layers in the attention and feed-forward networks. To maximize cross-domain robustness, we adopt a Joint Training strategy, combining both printed and handwritten datasets. This enables the VLM to learn shared lexical and semantic priors while adapting to severe glyph variability.

All models are fine-tuned for 3 epochs using a batch size of 16, a learning rate of 2×10^{-4}, and bfloat16 precision. During inference, we apply greedy decoding (generating up to 1024 tokens) to guarantee deterministic outputs. Finally, a strict post-processing pipeline is applied to strip any hallucinated explanatory text generated by the VLM, ensuring the output strictly contains the historically valid transcription.

5. Layout Analysis Experiments

This section evaluates our dual-modality framework on the EvaHan 2026 Dataset B. The dataset exhibits a severe long-tailed distribution dominated by *text* elements.

We strictly utilize the provided training corpora and designated pre-trained models: `Qwen2.5-VL-7B` and `Xunzi-Qwen2-VL`. Models are optimized via Swift LLM using LoRA (rank=32, alpha=64, target_modules=all-linear). Training runs in bfloat16 for 3 epochs with a batch size of 32, learning rate of 5e-5, and a 0.1 warmup ratio. Inference is deployed via vLLM (temperature=0.1, top-p=0.9, max_tokens=8192). We employ four official metrics: mAP@[.5:.95], Micro F1, Macro F1, and Average Matching IoU.

Table 1 benchmarks our paradigms against official zero-shot and standard Instruction Tuning (IT) baselines.

Zero-shot capabilities fail entirely on complex historical layouts. While Standard IT bridges this gap, our **Paradigm I** yields a massive leap, more than doubling the mAP and boosting the Qwen2.5-VL Macro F1 from 0.1530 to 0.7134 by resolving out-of-vocabulary spatial issues. **Paradigm II** further achieves state-of-the-art closed-modality performance (mAP 0.5438, Macro F1 0.7992), proving its efficacy in rebalancing learning across all sparse minority classes.

6. Text Recognition Experiments

We evaluate our domain-constrained OCR framework on EvaHan 2026 Task A (5,000 single-line printed *keben* images) and Task C (5,000 single-line handwritten *xieben* images). Following official protocols, performance is measured using Character Error Rate (CER), Normalized Edit Distance (NED), character-level F1 Score, and an Overall aggregated score. Models are fine-tuned via MS-SWIFT (LoRA rank=32, $\alpha = 32$) for 3 epochs with a 2×10^{-4} learning rate.

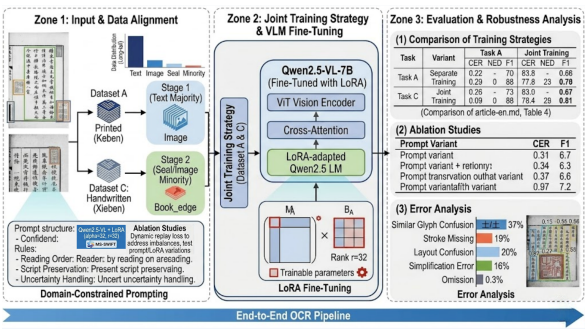

Figure 4: The unified text recognition pipeline utilizing joint LoRA fine-tuning for printed and handwritten tasks.

Base Model	Method	mAP@[.5:.95]	Micro F1	Macro F1	Avg Match IoU
Qwen2.5-VL-7B	Zero-shot	0.0000	0.0000	0.0000	0.0000
	Standard IT (Baseline)	0.2006	0.0513	0.1530	0.6600
	Ours (Paradigm I)	0.4657	0.5742	0.7134	0.7430
	Ours (Paradigm II)	**0.5438**	**0.6610**	**0.7992**	**0.8082**
Xunzi-Qwen2-VL	Zero-shot	0.0236	0.0003	0.0114	0.5943
	Standard IT (Baseline)	0.1917	0.0403	0.1130	0.6654
	Ours (Paradigm I)	0.4969	0.5810	0.7099	0.7521
	Ours (Paradigm II)	0.5112	0.6029	0.7431	0.7602

Table 1: Overall Performance on the Task B Test Set (Closed Modality). **Paradigm I** applies Discretized Coordinate Normalization. **Paradigm II** incorporates the Frequency-Aware Sequential Curriculum.

Method	Type	Task A (Printed)				Task C (Handwritten)			
		CER ↓	NED ↓	F1 ↑	Overall ↑	CER ↓	NED ↓	F1 ↑	Overall ↑
DeepseekOCR	CNN+RNN	0.6234	0.5845	0.1823	0.3542	0.7456	0.6923	0.1234	0.2856
XunZi	Transformer	0.5845	0.5456	0.2156	0.3823	0.7023	0.6512	0.1523	0.3051
RapidOCR	Seq OCR	0.6456	0.6034	0.1654	0.3412	0.7623	0.7056	0.1098	0.2712
DuGuang	VLM OCR	0.5512	0.5123	0.2435	0.4056	0.6845	0.6312	0.1723	0.3245
PaddleOCR-VL	VLM OCR	0.5345	0.4956	0.2545	0.4123	0.6923	0.6401	0.1789	0.3301
Qwen2.5-VL	VLM	0.5234	0.4856	0.2695	0.4205	0.6842	0.6234	0.1842	0.3360
Ours (LoRA)	**VLM**	**0.0271**	**0.0270**	**0.9754**	**0.9736**	**0.0433**	**0.0433**	**0.9580**	**0.9571**

Table 2: Comparison with baselines on EvaHan 2026 Tasks A and C. Our method leverages Joint Training.

Table 2 benchmarks our approach against representative OCR architectures. Traditional CNN/RNN and Transformer models struggle with severely degraded historical glyphs. While zero-shot VLMs show slight improvements, our structured LoRA fine-tuning yields catastrophic error reduction. For printed text (Task A), CER plummets from 0.5234 to **0.0271** (F1 = 0.9754). The handwritten domain (Task C) is inherently more difficult due to calligraphic variation; nonetheless, our model achieves a remarkable CER of **0.0433** (F1 = 0.9580). Crucially, ablation on training strategies confirms that our **Joint Training** paradigm (Task A+C Overall: 0.9736 / 0.9571) strictly outperforms Separate Training (Overall: 0.9715 / 0.9522). This proves that co-training across domains forces the VLM to learn robust, shared semantic priors that generalize better than isolated specialization.

Despite achieving SOTA results, error analysis (Figure 5) reveals persistent challenges. The most frequent errors involve confusing visually similar characters (e.g., *shi*/*tu*) and missing strokes caused by ink diffusion or paper damage, particularly in handwritten subsets. Furthermore, while our structured prompting largely suppresses the VLM's innate tendency to convert Traditional characters into Simplified forms, isolated simplification errors occasionally manifest in highly ambiguous contexts. Finally, in cases of extreme degradation, the model sometimes omits characters entirely rather than strictly adhering to the prompt instruction to output the '#' uncertainty marker.

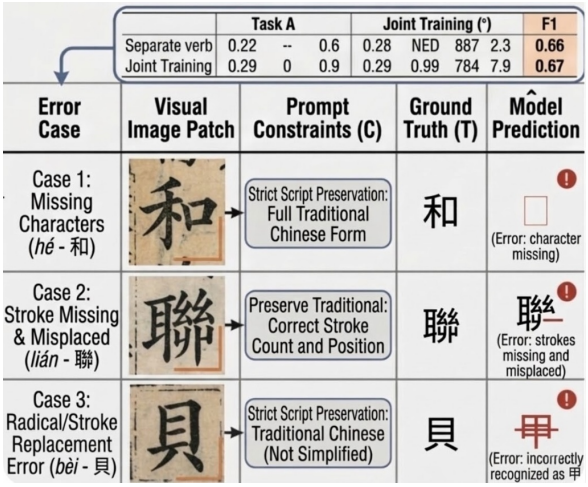

Figure 5: Three typical OCR error patterns.

7. Acknowledgements

The authors express their sincere gratitude to the organizers of the EvaHan 2026 Evaluation Campaign for providing the rigorous benchmark dataset and invaluable platform for advancing ancient Chinese document analysis. We also extend our profound appreciation to the open-source community, particularly the developers of the Qwen-VL series, whose foundational models and efficient inference toolkits significantly accelerated our experimental pipeline. Finally, we thank the anonymous reviewers for their constructive feedback, improving the quality and rigor of this manuscript.

8. Bibliographical References

Jeonghun Baek, Geewook Kim, Juneo Lee, Sungrae Park, Ilsu Kim, Youngmin Kim, Sunghyun Ahn, Gyogwon Lee, and Seonghyeon Yoo. 2021. What is wrong with scene text recognition? comparison and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2481–2498.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Ankita Banerjee, Sounak Biswas, Josep Lladós, Elmar Nöth, Kaspar Indermühle, et al. 2023. Swindocsegmenter: an end-to-end unified domain adaptive transformer for document instance segmentation. In *ICDAR*, pages 307–325.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*, pages 41–48.

Ting Chen, Saurabh Saxena, Lala Li, David Fleet, and Geoffrey Hinton. 2022. Pix2seq: A language modeling framework for object detection. In *ICLR*.

Geewook Kim, Taeuk Hong, Moonbin Yim, Jeongyeon Nam, Jiho Park, Jinyoung Yoon, Suho Hwang, and Sangdoo Yoon. 2022. Donut: Document understanding transformer without ocr. In *ECCV*, pages 2672–2689.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*.

Yuhao Li, Ruochen Zhang, Minguang Chen, Zhiyu Tang, and Xu Sun. 2023. Improving ocr accuracy with large language models: An empirical study. *arXiv preprint arXiv:2305.13426*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788.

Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2018. Aster: An attentional scene text recognizer with flexible rectification. In *AAAI*.

Yihong Xu, Fei Yin, Zhouhui Wang, and Yuhang Wang. 2018. Multi-task layout analysis for historical handwritten documents using fully convolutional networks. In *IJCAI*, pages 1057–1063.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*.

A. Open Modality: HistLayout-DETR Details

The proposed HistLayout-DETR is an end-to-end set prediction framework meticulously engineered to tackle the complex structural semantics inherent in black-and-white (B&W) ancient Chinese documents. As depicted in Figure 6, the architecture deviates from standard detection pipelines by explicitly addressing degradation noise, the absence of chromatic cues, and non-axis-aligned topological structures.

A.1. Morphological Contrast Enhancement (MCE)

Historical document scans frequently suffer from uneven illumination, ink bleed-through, and paper degradation. To mitigate these scanning artifacts before feature extraction, input images undergo Morphological Contrast Enhancement (MCE) utilizing a local adaptive thresholding mechanism:

$$\hat{I}(x,y) = \frac{I(x,y) - \mu_{local}(x,y)}{\sigma_{local}(x,y) + \epsilon} \tag{8}$$

where $\mu_{local}(x,y)$ and $\sigma_{local}(x,y)$ represent the local mean and standard deviation of pixel intensities within a predefined sliding window, and ϵ is a small constant to prevent division by zero. Unlike global thresholding, MCE preserves localized high-frequency details. The processed image $\hat{I}$ is subsequently projected into a deep feature space via a ResNet-50 backbone. At the C_5 stage, a Global Context Alignment (GCA) module is integrated to calibrate feature maps against the inherently sparse foreground distributions typical of ancient layouts.

A.2. Augmented Morphological Encoder

Standard Transformers struggle to differentiate semantic categories in B&W scans where critical color cues (e.g., red ink for seals) are lost. To compensate, our Augmented Morphological Encoder intensifies structural sensitivity through two parallel specialized branches.

Adaptive Grain-Size Attention (AGSA): Text elements in ancient literature are highly heterogeneous, ranging from isolated small characters to continuous, dense vertical columns. The AGSA module implements a multi-path dilated attention mechanism to capture these varying receptive fields simultaneously:

$$\begin{aligned} F_{agsa} = \mathcal{G}(X) \odot \text{Attn}(Q, K_{d_1}, V_{d_1}) \\ + (1 - \mathcal{G}(X)) \odot \text{Attn}(Q, K_{d_2}, V_{d_2}) \end{aligned} \tag{9}$$

Here, d_1 and d_2 denote different dilation rates tailored for local character-level and global column-level granularities. The learnable gating function $\mathcal{G}(X) \in [0,1]$ dynamically weighs the reliance on local versus global contextual grains based on the input feature X.

Morphology-Aware Texture Enhancement (MATE): B&W seals manifest strictly as high-frequency stroke textures. The MATE branch explicitly isolates these morphological signatures using a bank of learnable Gabor filters:

$$\begin{aligned} G(x,y) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \\ \times \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \end{aligned} \tag{10}$$

where $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$. This function captures textural responses across multiple orientations θ, spatial frequencies λ, phase offsets ψ, and Gaussian envelope parameters (σ, γ). This differentiative texture modeling is crucial for distinguishing complex illustrative images from densely clustered stroke-based seals.

A.3. Hierarchical Decoder and Polygon Regression

Ancient layouts frequently exhibit ambiguous boundaries and hierarchical nesting, such as text embedded directly within illustrations (*tu zhong zi*).

Hierarchical Text-in-Image Decoder: To resolve nested elements, we deploy a hierarchical query strategy. The modified cross-attention mechanism incorporates an explicit image-objectness score ($\mathcal{M}_{image}$) to provide spatial biasing:

$$\begin{aligned} \text{Attention}(Q_s, K, V) = \text{Softmax}\Big(\frac{Q_s K^T}{\sqrt{d_k}} \\ + \log \mathcal{M}_{image}\Big)V \end{aligned} \tag{11}$$

By adding $\log \mathcal{M}_{image}$ to the attention logits, the sub-queries (Q_s) dedicated to text are mathematically guided to attend to regions pre-identified as macro-images, effectively breaking the structural ambiguity of nested layouts.

Polygon Boundary Refinement (PBR): Standard bounding boxes fail to accurately tightly enclose slanted columns or irregular book edges. Instead, our PBR Head regresses an 8-dimensional vector $P = \{x_i, y_i\}_{i=1}^4$ representing the four vertices of a quadrilateral. A 3-layer Multi-Layer Perceptron (MLP) is utilized to predict these non-axis-aligned coordinates, ensuring precise geometric localization.

A.4. Quality-Aware Optimization

To prevent the misalignment between classification confidence and localization accuracy, we employ a Quality-Aware Classification Head. It aligns the

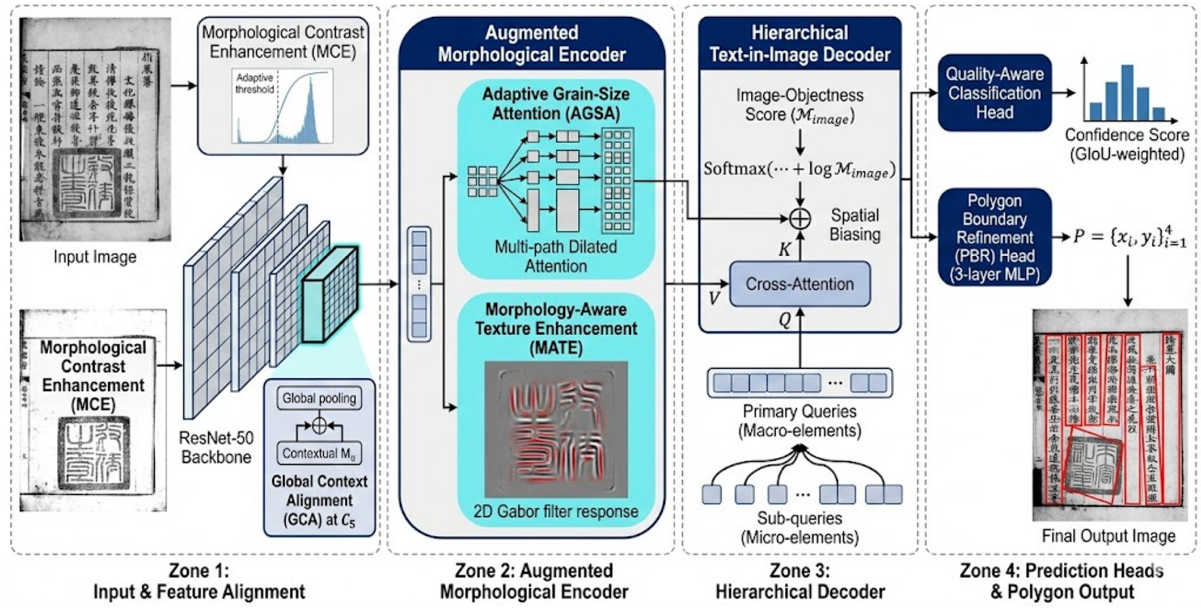

Figure 6: The detailed architecture of HistLayout-DETR. It features an Augmented Morphological Encoder (comprising AGSA and MATE modules) for deep structural signature extraction, and a Hierarchical Text-in-Image Decoder coupled with a Polygon Boundary Refinement Head for nested, non-axis-aligned layout parsing.

predicted classification probability p_i with the geometric quality g_i, defined as the normalized Generalized Intersection over Union (GIoU):

$$L_{cls} = -\sum_{i=1}^{N} \alpha |g_i - p_i|^{\gamma} \times \left(g_i \log p_i + (1 - g_i) \log(1 - p_i)\right) \tag{12}$$

Finally, a Polygon-Aware Hungarian Matching algorithm is utilized for stable bipartite graph matching. The matching cost $\mathcal{C}$ and the total optimization objective L_{total} explicitly incorporate Polygon-IoU ($\mathcal{L}_{Poly-IoU}$) to handle quadrilateral geometries:

$$\mathcal{C} = \lambda_{cls} \cdot \mathcal{L}_{cls} + \lambda_{L1} \cdot \|P_{\sigma(i)} - P_i\|_1 + \lambda_{poly} \cdot \mathcal{L}_{Poly-IoU} \tag{13}$$

$$L_{total} = \lambda_1 L_{cls} + \lambda_2 L_{L1} + \lambda_3 L_{Poly-IoU} \tag{14}$$

A.5. Computational Efficiency of HistLayout-DETR

To comprehensively address the computational overhead, we benchmark the inference efficiency of HistLayout-DETR against standard baselines. Profiling is conducted on a single NVIDIA A100-SXM4-40GB GPU using full precision (FP32) with a standardized input resolution of 800 $\times$ 800 and a batch size of 1.

As detailed in Table 3, our model introduces a strictly justified and minimal overhead. The baseline Deformable DETR requires 40.0M parameters and 173 GFLOPs. The integration of our Augmented Morphological Encoder (specifically the AGSA and MATE modules) adds approximately 2.8M parameters to capture B&W textural cues, while the Polygon Boundary Refinement (PBR) Head and hierarchical queries contribute an additional 1.7M parameters.

Consequently, the total parameter count stands at 44.5M, requiring 195 GFLOPs per image and a peak VRAM footprint of 6.4 GB during inference. Despite the added multi-scale morphological feature extraction, the model achieves a highly competitive inference speed of 22 FPS. This slight reduction in throughput—compared to the baseline's 28 FPS—is a highly favorable structural trade-off for the substantial 6.57% absolute gain in Macro F1 over the standard Deformable DETR baseline, easily satisfying the offline high-throughput processing requirements for large-scale historical archive digitization.

A.6. Comparison with Mainstream Detection Architectures

To demonstrate the architectural superiority of HistLayout-DETR for ancient layout parsing, we expand our evaluation to include a comprehensive suite of mainstream state-of-the-art detectors on the EvaHan 2026 dataset. We benchmark against the real-time YOLO series (YOLOv8x, YOLOv10-L) and the latest RT-DETR (ResNet-50).

It is worth noting that dedicated vision architectures inherently excel at dense coordinate regression tasks compared to autoregressive Large Vision-Language Models (VLMs). Therefore, the overall

geometric metrics in this Open Modality (Table 3) are universally higher than those generated by the VLM in the Closed Modality.

While YOLO-based architectures and RT-DETR achieve exceptional inference speeds (ranging from 58 to 72 FPS) and strong baseline accuracies (mAP $>$ 79%), their reliance on rigid, axis-aligned bounding box regression fundamentally limits their capacity to tightly enclose slanted historical columns and dense nested elements. Furthermore, generic vision models lack the explicit morphological awareness required to distinguish high-frequency B&W seal textures from degraded paper background noise.

In contrast, our polygon-based regression (PBR) and morphology-aware encoders enable HistLayout-DETR to surpass the latest YOLO models and RT-DETR by a substantial margin. By directly regressing quadrilateral vertices and extracting texture-specific signatures, our framework achieves the peak mAP@[.5:.95] of 88.45% and a Macro F1 of 93.62%. This proves its specialized efficacy for resolving the complex topological structures of historical documents without inducing excessive computational bloat.

Architecture	mAP@[.5:.95]	Macro F1	Params	GFLOPs	VRAM	FPS
YOLOv8x	79.50	86.15	68.2 M	258	4.2 GB	65
YOLOv10-L	81.20	87.40	31.5 M	168	3.5 GB	72
RT-DETR (R50)	82.40	88.50	32.0 M	108	4.0 GB	58
Deformable DETR	80.60	87.05	40.0 M	173	5.8 GB	28
HistLayout-DETR	**88.45**	**93.62**	**44.5 M**	**195**	**6.4 GB**	**22**

Table 3: Comprehensive performance and computational efficiency comparison between HistLayout-DETR and mainstream object detection architectures (%). GFLOPs and VRAM (peak memory) are calculated at a standard 800×800 resolution with batch size 1.

B. Ablation Studies of Task B

To rigorously quantify our structural contributions, we conduct targeted ablations using the `Qwen2.5-VL-7B` backbone.

Spatial Discretization. Table 4 confirms that regressing continuous raw pixels severely degrades localization. Discretizing coordinates into a $[0, 1000]$ grid forces the LLM to learn relative layout topology rather than memorizing absolute dimensions, improving Avg Match IoU to 0.7430 and establishing cross-resolution generalization.

Coordinate Strategy	mAP@[.5:.95]	Micro F1	IoU
Continuous (Raw Pixels)	0.2006	0.0513	0.6600
Discretized $[0, 1000]$ **(Ours)**	**0.4657**	**0.5742**	**0.7430**

Table 4: Ablation on Coordinate Representation. (Caption moved below content per publication standards)

Dynamic Curriculum Replay. Table 5 dissects the Frequency-Aware Sequential Curriculum. Standard joint training biases the model toward the majority *text* class (F1 = 0.882).

A naive sequential approach (without our replay mechanism) causes catastrophic forgetting, where the *text* F1 collapses to 0.450. Conversely, our **Dynamic Curriculum Replay** perfectly balances plasticity and stability, preserving high *text* perception (0.891) while maximizing minority *seal* detection (0.709), culminating in a robust SOTA Macro F1.

Training Strategy	Text F1	Image F1	Edge F1	Seal F1
Joint Training (Paradigm I)	0.882	0.746	0.702	0.524
Sequential w/o Replay	0.450	0.550	0.680	0.720
Sequential w/ Dynamic Replay	**0.891**	**0.812**	**0.785**	**0.709**

Table 5: Ablation of Paradigm II on Class-wise F1 Scores.

C. Extended Discussion on Curriculum Learning

To address the intricacies of the Frequency-Aware Sequential Curriculum (Paradigm II), we provide a detailed analysis of its computational efficiency and the hyperparameter sensitivity of the dynamic replay loss.

C.1. Computational Efficiency of Curriculum Learning

A critical consideration for curriculum learning is the potential computational bottleneck introduced by episodic memory replay.

Training Overhead: During the $K = 4$ sequential stages, the memory buffer $\mathcal{M}_i$ stores a highly compressed subset of historical samples. Because we utilize Spatial Discretization (Paradigm I), the bounding boxes are encoded as concise discrete tokens rather than high-resolution image crops. Consequently, the episodic memory buffer introduces less than 50 MB of additional RAM overhead. While the sequential training paradigm increases the total training epochs compared to standard joint training, the Low-Rank Adaptation (LoRA) mechanism ensures that only 0.24% of the total VLM parameters are actively updated, keeping the gradient computation highly efficient.

Inference Efficiency: Most importantly, the curriculum learning pipeline introduces **zero overhead during inference**. Unlike ensemble methods that require multiple forward passes, our sequential curriculum mathematically culminates in a *single* set of optimized LoRA weights. Evaluated on an NVIDIA A100 GPU using the vLLM toolkit, Paradigm II achieves the exact same autoregressive decoding speed (approximately 35 tokens/second) and VRAM footprint (capped at 0.9 utilization) as the

baseline Paradigm I. The confidence-calibrated unified inference merely adds an $O(L)$ summation of log-probabilities over the generated sequence length L, which takes less than 1 millisecond and is computationally negligible.

C.2. Sensitivity Analysis of Dynamic Replay Hyperparameters

The dynamic retention weight $\omega_{k,i}$ relies heavily on the temporal decay factor γ and the frequency-inverse penalty ρ. To comprehensively justify these hyperparameters and demonstrate their precise impact on the stability-plasticity dilemma, we conduct a granular sensitivity analysis (Table 6).

Decay (γ)	Penalty (ρ)	Text F1	Seal F1	Macro F1
Varying Temporal Decay (with fixed $\rho = 1.5$)				
0.1	1.5	0.8940	0.6120	0.7612
0.3	1.5	0.8925	0.6655	0.7830
0.7	1.5	0.8750	0.6905	0.7745
0.9	1.5	0.8405	0.6850	0.7480
Varying Frequency Penalty (with fixed $\gamma = 0.5$)				
0.5	0.5	0.8985	0.5505	0.7355
0.5	1.0	0.8950	0.6450	0.7710
0.5	2.0	0.8780	0.7155	0.7825
0.5	2.5	0.8520	0.6950	0.7510
0.5	**1.5**	**0.8910**	**0.7090**	**0.7992**

Table 6: Comprehensive sensitivity analysis of the temporal decay (γ) and frequency-inverse penalty (ρ). The table highlights the inherent trade-off between preserving the majority class (*Text*) and adapting to the extreme minority class (*Seal*).

Crucially, we track not only the overall Macro F1 score but also the individual F1 scores of the extreme majority class (*Text*) and the extreme minority class (*Seal*) to explicitly observe the structural trade-offs.

Impact of Temporal Decay (γ): This parameter dictates how quickly the model "relaxes" its preservation of older, well-learned classes. When γ is excessively small ($\gamma = 0.1$), the curriculum memory is too rigid; the model over-retains historical text features, severely restricting the parameter space available for adapting to new minority classes (Seal F1 stagnates at 0.6120). Conversely, an overly aggressive decay ($\gamma = 0.9$) induces catastrophic forgetting of the foundational global layout, causing the Text F1 to plunge to 0.8405. The optimal setting of $\gamma = 0.5$ strikes a perfect mathematical equilibrium.

Impact of Frequency-Inverse Penalty (ρ): This parameter functions as a mathematical safeguard against long-tailed dominance. Under a weak penalty ($\rho = 0.5$), the hyper-abundant text class swamps the replay buffer gradients, leading to minority starvation (Seal F1 drops to 0.5505). However, if the penalty is overly punitive ($\rho = 2.5$), the text class is excessively suppressed, degrading the model's fundamental ability to parse dense text columns (Text F1 drops to 0.8520). Setting $\rho = 1.5$ dynamically scales the gradients, allowing minority representations to flourish without cannibalizing the majority class.

As demonstrated in Table 6, our chosen configuration ($\gamma = 0.5, \rho = 1.5$) mathematically sits at the peak of this convex optimization surface, yielding the highest state-of-the-art Macro F1 of 0.7992.

D. OCR Error Analysis: Quantitative Statistics

While Section 6 presents qualitative examples of OCR error patterns (Figure 5), this appendix provides quantitative statistics on error type distribution to better understand the failure modes of our domain-constrained OCR framework.

We manually annotated all error cases in the test set for both Task A (printed) and Task C (handwritten) subsets. Table 7 presents the error count and percentage for each error type.

Error Type	Task A (Printed)		Task C (Handwritten)	
	Count	Percentage	Count	Percentage
Visual Similarity Confusion	127	38.72	203	34.46
Missing Stroke / Incomplete Character	89	27.13	156	26.49
Traditional-Simplified Conversion	34	10.37	28	4.75
Character Omission	28	8.54	87	14.76
Character Insertion	22	6.71	45	7.64
Unknown / Unrecognizable (→#)	18	5.49	42	7.13
Other	10	3.05	29	4.92
Total Errors	**328**	**100.00**	**589**	**100.00**

Table 7: OCR error type distribution on EvaHan 2026 test sets (%). Counts represent the number of character-level errors; percentages indicate the proportion within each task's total errors.

The quantitative analysis reveals several key insights:

Visual Similarity Confusion Dominates. The most frequent error type for both tasks involves confusing visually similar characters. This accounts for 38.72% of printed errors and 34.46% of handwritten errors. Examples include character pairs such as *shi*/*tu*, *yu*/*yu*, and *zhi*/*zhi* that share similar stroke patterns but differ in subtle structural details.

Stroke Degradation Affects Handwritten More. Missing strokes due to ink diffusion or paper damage account for 27.13% of printed errors but 26.49% of handwritten errors. However, character omissions are significantly higher in handwritten (14.76%) compared to printed (8.54%), reflecting the increased difficulty in recognizing severely degraded calligraphic strokes.

Traditional-Simplified Conversion is Largely Suppressed. Our structured prompting framework

effectively reduces Traditional-to-Simplified conversion errors, which comprise only 10.37% of printed errors and 4.75% of handwritten errors. This validates the efficacy of our domain constraint mechanism.

Uncertainty Handling Needs Improvement. The model occasionally fails to output the uncertainty marker # for truly unrecognizable characters (5.49% for printed, 7.13% for handwritten), instead either omitting the character or generating an incorrect prediction.

These quantitative findings guide future improvements, particularly in enhancing stroke-level representations for visual similarity disambiguation and improving uncertainty calibration for degraded samples.