

LadderMoE: Ladder-Side Mixture of Experts Adapters for Bronze Inscription Recognition

Rixin Zhou¹, Peiqiang Qiu², Qian Zhang², Chuntao Li^{2,3*},
Xi Yang^{1,3,4*}

¹School of Artificial Intelligence, Jilin University.

²School of Archaeology, Jilin University.

³Key Laboratory of Ancient Chinese Script, Culture Relics and
Artificial Intelligence, Jilin University.

⁴Engineering Research Center of Knowledge-Driven Human-Machine
Intelligence, MoE.

*Corresponding author(s). E-mail(s): lct33@jlu.edu.cn;
yangxi21@jlu.edu.cn;

Abstract

Bronze inscriptions (BI), engraved on ritual vessels, constitute a crucial stage of early Chinese writing and provide indispensable evidence for archaeological and historical studies. However, automatic BI recognition remains difficult due to severe visual degradation, multi-domain variability across photographs, rubbings, and tracings, and an extremely long-tailed character distribution. To address these challenges, we curate a large-scale BI dataset comprising 22,454 full-page images and 198,598 annotated characters spanning 6,658 unique categories, enabling robust cross-domain evaluation. Building on this resource, we develop a two-stage detection–recognition pipeline that first localizes inscriptions and then transcribes individual characters. To handle heterogeneous domains and rare classes, we equip the pipeline with LadderMoE, which augments a pretrained CLIP encoder with ladder-style MoE adapters, enabling dynamic expert specialization and stronger robustness. Comprehensive experiments on single-character and full-page recognition tasks demonstrate that our method substantially outperforms state-of-the-art scene text recognition baselines, achieving superior accuracy across head, mid, and tail categories as well as all acquisition modalities. These results establish a strong foundation for bronze inscription recognition and downstream archaeological analysis.

Keywords: BI Recognition, Mixture-of-Experts, Parameter-Efficient Fine-Tuning

1 Introduction

Bronze inscriptions (BI), engraved on ritual vessels of ancient China, constitute a crucial component of the early Chinese writing system alongside oracle bone inscriptions (OBI), preserving invaluable records of early civilization [1]. Western Zhou inscriptions, for example, document royal rewards, sacrificial rituals, military campaigns, and political appointments [2]. Figure 1 (A) illustrates representative BI data across three typical forms: color photographs, rubbings, and tracings. Accurate recognition of such heterogeneous inscriptions is vital for downstream applications including bronze dating, archaeogeographical analysis, and historical literature retrieval, providing a reproducible bridge from raw imagery to cultural-heritage research, as shown in Figure 1 (D).

Traditionally, the study of bronze inscriptions has relied on manual rubbings, tracings, and philological analysis, a process that is labor-intensive and heavily dependent on expert knowledge. With the rapid progress of computer vision, automatic detection and recognition of ancient scripts has emerged as a promising alternative. However, BI recognition remains highly challenging (Figure 1(B)) due to multi-domain diversity (color photographs, rubbings, and tracings), pronounced degradation/noise and frequent low resolution from centuries of weathering and uneven casting, and a severe long-tailed character distribution in which common ritual or administrative symbols dominate while personal names, clan titles, and toponyms are intrinsically rare [3]. These factors impede the direct transfer of methods developed for OBI or modern scene text.

Prior work has largely centered on OBI, exploring improved detectors and glyph-structure-guided methods [4–6], with only limited extensions to BI [7, 8] that typically rely on heavy preprocessing and rubbings, leaving real-scene photographs underexplored. Meanwhile, transformer-based scene text recognition methods show promise [9–11], but its context-aware language priors are unreliable for BI because the specialized vocabulary is scarcely represented in large pretraining corpora.

To overcome the key challenges of bronze inscription recognition and the limitations of existing research, we present the following contributions:

- We curate a large-scale bronze inscription dataset comprising 22,454 full-page images with 198,598 annotated character across 6,658 unique categories, spanning color photographs, rubbings, and tracings to support robust cross-domain evaluation.
- We build a two-stage pipeline for full-page BI recognition that first detects inscriptions and then performs character recognition and transcription (Figure 1 C). Within this framework, we propose the LadderMoE, a parameter-efficient model based on a pretrained CLIP image encoder that interleaves lightweight experts across multiple transformer layers, enabling efficient training and expert specialization to handle domain heterogeneity and rare-class patterns.
- Comprehensive experiments demonstrate that our framework surpasses existing methods in overcoming the key challenges of multi-domain variation, visual degradation, and long-tailed character distribution, and achieves state-of-the-art

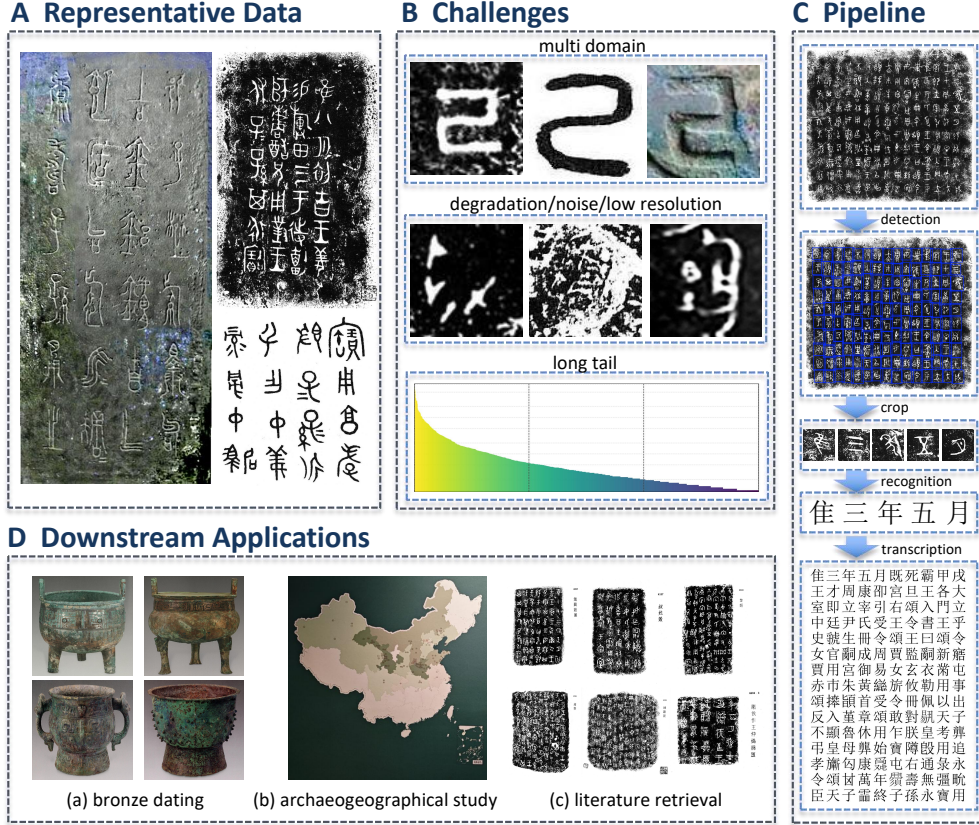


Fig. 1 Overview of our problem setting and approach for full-page bronze inscription recognition. The illustration emphasizes the cross-domain, degraded, and long-tailed nature of the data and motivates a detection–recognition–ordering framework that is robust to these factors. The resulting structured transcriptions enable downstream archaeological analyses, including bronze dating, archaeogeographical study, and literature retrieval.

performance on both single-character and full-page bronze inscription recognition tasks.

2 Related Work

2.1 Ancient Chinese Inscription Detection and Recognition

Research on ancient Chinese script recognition has long centered on Oracle Bone Inscriptions (OBI), while Bronze Inscriptions (BI) remain comparatively underexplored. Early detection studies simply adapted generic object detectors. For example, Liu et al. enhanced Faster R-CNN for OBI character detection [4], Fu et al. introduced pseudo-category labels and glyph-structure priors to improve noise robustness [5], and Tao et al. leveraged the OBC font library with clustering-based representation learning for stronger feature extraction [6]. These methods established a foundation but

still inherit the limitations of generic detectors, including heavy dependence on pre-processing and insufficient adaptability to heterogeneous visual domains. Recognition techniques have progressed along two principal directions. Structure-driven pipelines extract line or stroke-level geometry and then perform geometric matching, e.g., via Hough transforms [12]. Such approaches explicitly encode stroke topology but are sensitive to background clutter and low-contrast corrosion. Learning-based models embed character images into discriminative feature spaces for nearest-neighbor or sequence matching [13, 14], and recent transformer variants—such as the improved Swin-Transformer [7] with pruning-based acceleration [15]—have further advanced recognition across OBI, BI, and stone engravings. Despite these advances, systematic BI detection and recognition remain largely unexplored. Existing pipelines generally assume well-preprocessed rubbings or clean tracings, which limits their robustness to real-scene photographs that exhibit complex casting textures, multi-domain variation (color photos, rubbings, and tracings), and significant visual degradation. Furthermore, the intrinsically long-tailed character distribution in bronze inscriptions poses challenges for balanced learning and evaluation.

2.2 Scene Text Recognition

Scene Text Recognition (STR) aims to read text from cropped regions in natural images and has enabled applications such as understanding road signs, product labels, and document analysis [9]. Unlike conventional OCR, STR must handle heterogeneous fonts, arbitrary orientations, curved layouts, and complex illumination, making it a particularly challenging problem. Recent progress has been driven by transformer-based sequence models [9, 16] and semi-supervised paradigms that exploit unlabeled data [17, 18], complementing earlier end-to-end architectures [19–21]. Methodologically, STR approaches fall into two categories. Context-free methods rely solely on visual evidence, including CTC-based recognizers [22–25], segmentation-driven pipelines [26, 27], and attention-based encoder–decoder models [28, 29]. Context-aware methods augment vision with linguistic priors, as in ABINet [10], CLIP-OCR [30], and CLIP4STR [11], which leverage external language models or cross-modal knowledge. The challenges faced in bronze inscription recognition closely parallel those of STR: characters appear on complex, uneven surfaces with variable lighting, occlusion, and background noise. Context-free STR techniques—which focus purely on robust visual modeling—provide a suitable foundation for recognizing BI from both rubbings and real-scene photographs.

2.3 Parameter-efficient Fine-tuning

Parameter-efficient fine-tuning (PEFT) adapts large pre-trained models to downstream tasks by updating only a small subset of parameters, thereby avoiding the computational and energy costs of full fine-tuning [31]. Representative PEFT families differ in where and what they tune: adapter tuning inserts lightweight bottleneck modules into Transformer layers [32, 33]; LoRA injects trainable low-rank matrices into frozen weight paths [34]; and prompt tuning optimizes task-specific, learnable prompts while keeping backbone weights fixed [35]. Orthogonal to PEFT, Mixture-of-Experts

(MoE) architectures expand model capacity via multiple experts and a routing network that sparsely activates only a small subset per input, enabling near-constant per-token compute while scaling representational power [36–39]. Although prior work has extensively studied PEFT and MoE in isolation, their combination is particularly appealing for domains with strong intra-class variability and modality/style heterogeneity—such as ancient script recognition—where efficient specialization and targeted parameterization are both desirable. Our work situates itself at this intersection by introducing ladder-side MoE-Adapters, which attach adapter experts along the backbone and employ routing to learn complementary representations for different types of BI. This design couples PEFT’s low-overhead adaptation with MoE’s selective expert allocation, yielding a parameter-efficient yet specialization-aware approach to bronze inscription recognition.

3 Methodology

3.1 Full-page Bronze Inscriptions Recognition Pipeline

We adopt a two-stage detect–then–recognize pipeline for full-page BI recognition, as shown in Figure 2 (a). An off-the-shelf object detector, YOLO-v12 [40], is first applied to full-page inscription images to localize character instances. The detected regions are then cropped into single-character patches and recognized by our LadderMoE. During training, the detector is learned on full-page images with bounding-box annotations, and the recognizer is trained on single-character crops generated from ground-truth boxes.

3.2 LadderMoE

3.2.1 Encoder

As illustrated in Figure 2 (b), we employ a pretrained CLIP image encoder, and MoE-Adapters are inserted at multiple intermediate layers through ladder-style connections. These adapters are governed by a unified router that dynamically selects a sparse subset of experts, enabling adaptive routing of features across categories with diverse characteristics. The outputs from the selected experts are combined and progressively fused with the backbone stream by a trainable gate, which is subsequently fed into an image decoder for final character code prediction.

3.2.2 Ladder-side MoE Adapter

Each MoE adapter contains a unified router responsible for selecting a sparse subset of experts from a pool of N candidate experts. Given an adapter input, the router first aggregates information from the class token and the average-pooled image token to form its routing signal. This signal is projected into a one-dimensional vector of expert scores, after which only the top- k experts with the highest scores are activated. The router then applies a softmax function to these selected scores to obtain normalized routing weights. Using these weights, the adapter computes a weighted sum of the outputs of the chosen experts, producing the final expert-enhanced representation.

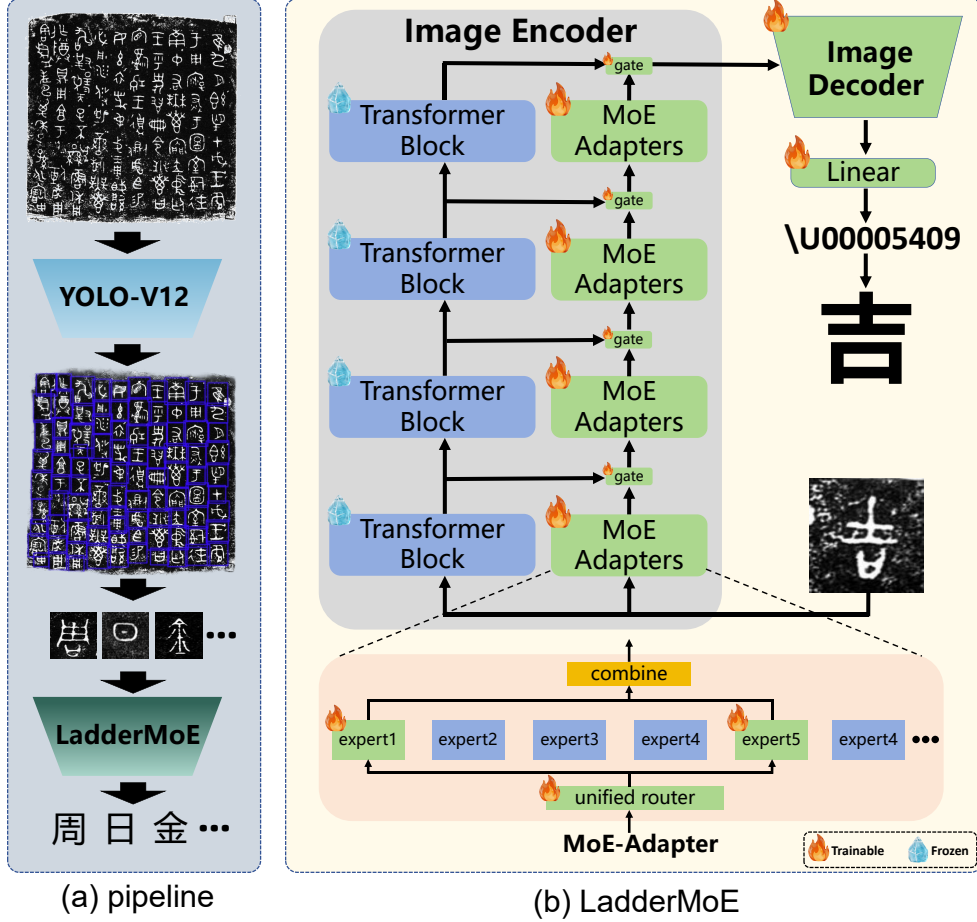


Fig. 2 Framework of our bronze inscriptions recognition model. A Transformer-based encoder is augmented with interleaved MoE-Adapters, and the enriched features are decoded into character code. Each MoE-Adapter consists of multiple experts and a unified router that dynamically selects a sparse subset of experts for the input.

3.2.3 Decoder

We adopt the same decoder architecture as PARSeq [9], which employs a shallow single-layer decoder to extract character information from the visual feature. Unlike PARSeq, which relies on Permutation Language Modeling (PLM) for training, we further introduce an Ordered Sequence Fine-tuning (OSF) stage. The character order of BI carries intrinsic semantic meaning. Therefore, during the later phase of training we replace the random attention masks used in PLM with a fixed sequential mask. The OSF stage strengthens the alignment between the predicted character sequence and its underlying semantic structure.

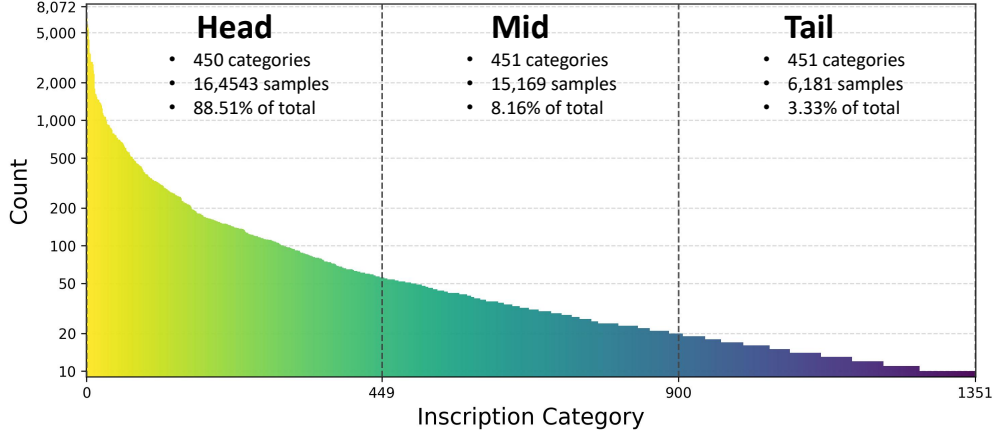


Fig. 3 Category frequency distribution of the 1,352 selected bronze inscription categories. The distribution exhibits a pronounced long-tailed pattern: a few characters occur thousands of times (up to over 8,000 instances), while the majority of classes appear only sparsely. This imbalance poses a significant challenge for recognition models and motivates our evaluation across Head, Mid, and Tail subsets.

4 Datasets

4.1 Data Filtering

We construct a large-scale dataset of BI comprising 22,454 images covering 6,658 distinct character categories. The most frequent character appears 8,072 times, whereas some characters occur only once, reflecting the extreme sparsity. Among the collected data, 3,037 are color photographs, while the remaining 17,360 are rubbings and 2070 are tracings, capturing both archaeological records and traditional research materials. To ensure that the detection and recognition networks are trained on categories with sufficient visual evidence, we retain only those categories with more than 10 samples, discarding extremely sparse categories. The final filtered dataset contains 17,002 images across 1,352 inscription categories, and all subsequent bronze-inscription detection and recognition experiments are conducted on this refined subset.

4.2 Data Splits

For the full-page inscription detection and recognition task, we further split 17,002 filtered images into training, validation, and test sets with an 8 : 1 : 1 ratio, ensuring that each split preserves the same distribution of color, rubbing, and tracing images to maintain domain consistency across splits.

To evaluate the single-char recognition task alone, we crop individual characters of the 1,352 categories from the original images, resulting in 185,893 character patches. Each character category is then divided into training, validation, and test subsets with a 4 : 1 : 5 ratio, guaranteeing balanced coverage of every category.

To enable a more comprehensive evaluation of recognition methods under category imbalance, we divide the 1,352 character categories into three groups—Head, Mid,

and Tail—based on frequency, ensuring that each group contains approximately one-third of the categories, as shown in Figure 3. This stratification allows us to analyze model performance across characters with abundant, moderate, and scarce training examples, providing insights into robustness under real-world long-tailed scenarios.

5 Experiments

5.1 Implementation Details

5.1.1 Devices and Code.

All experiments were implemented by PyTorch, and conducted on a server with 4 RTX A40 GPUs and Intel® Xeon® Gold 5220 CPUs (72 cores). For fair comparison, we adopt the official implementations of all baseline methods.

5.1.2 Training Details.

We set the batch size to 32 and train models for 40 epochs in total. Specifically, the first 35 epochs use permuted sequence masks to encourage diverse dependency learning, followed by 5 epochs ordered sequence fine-tuning, and the number of permutations for sequence modeling is set to 12. In the MoE modules, we use 36 experts per layer with top-5 expert selection.

To reduce training cost, MoE-Adapters are placed only at selected encoder layers [0, 4, 8, 11]. During training, the backbone encoder parameters are frozen, while the learnable gate, unified router, activated experts and the decoder remain learnable.

5.1.3 Transcription Algorithm.

At inference, the detected boxes are first passed to the recognition model to obtain character predictions, after which Algorithm 1 adaptively estimates a horizontal threshold and clusters the boxes into right-to-left columns with top-to-bottom ordering, producing a structured full-page transcription result.

5.1.4 Evaluation Metric.

Single-Character Recognition. We evaluate single-inscription recognition using multiple accuracy measures to assess overall performance and robustness across class imbalance and domain shifts. Let the test set be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where x_i is the input and y_i the ground-truth label. Denote the set of all classes as with $\mathcal{C}$ cardinality $|\mathcal{C}|$. For each class $c \in \mathcal{C}$, let $\mathcal{D}_c = \{i \mid y_i = c\}$ be the index set of its samples, and let $\hat{y}_i$ be the predicted label for sample i . The indicator function $\mathbf{1}[\cdot]$ equals 1 if the condition inside is true and 0 otherwise. The overall accuracy is defined as:

$$\text{Overall Acc} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{y}_i = y_i], \quad (1)$$

Algorithm 1: Column-wise Grouping for Full-page Transcription

Input: Detected text boxes $\mathcal{B} = \{b_i\}$, each $b = ((x_1, y_1), (x_2, y_2))$; scaling factor λ (default 0.5).

Output: Ordered columns $\mathcal{C} = [C_1, \dots, C_M]$.

Function `ComputeAdaptiveThreshold`($\mathcal{B}$, $factor$):

```
     $W \leftarrow \{x_2 - x_1 \mid b = ((x_1, y_1), (x_2, y_2)) \in \mathcal{B}, x_2 > x_1\};$   
     $\bar{w} \leftarrow \frac{1}{|W|} \sum_{w \in W} w;$   
    return  $\bar{w} \times \lambda;$ 
```

Main Procedure:

$x_{\text{thr}} \leftarrow \text{ComputeAdaptiveThreshold}(\mathcal{B}, factor);$

Sort $\mathcal{B}$ by x_1 in *descending* order (rightmost first);

$\mathcal{C} \leftarrow [];$

foreach $b \in \mathcal{B}$ **do**

 assigned $\leftarrow$ **false**;

foreach $C \in \mathcal{C}$ **do**

$x^{\text{anchor}} \leftarrow x_1$ of the first box in C ;

if $|x_1(b) - x^{\text{anchor}}| < x_{\text{thr}}$ **then**

 append b to C ; assigned $\leftarrow$ **true**; **break**;

if not assigned then

 create new column $C^* \leftarrow [b]$ and append to $\mathcal{C}$;

foreach $C \in \mathcal{C}$ **do**

 sort C by y_1 in ascending order (top $\rightarrow$ bottom)

Sort $\mathcal{C}$ by x_1 of the first box in *descending* order (right $\rightarrow$ left);

return $\mathcal{C}$;

The class-balanced average accuracy is defined as:

$$\text{Balanced Acc} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} \mathbf{1}[\hat{y}_i = y_i]. \quad (2)$$

To evaluate robustness across class-frequency regimes and acquisition domains, we further report accuracies on specific subsets of the test data. Let $\mathcal{D}_H$, $\mathcal{D}_M$, and $\mathcal{D}_T$ denote the sample indices belonging to head, mid, and tail classes respectively. Similarly, let $\mathcal{D}_d$ represent samples from a particular domain d (e.g., color, rubbing and tracing images). The accuracy on any subset $S \subseteq \mathcal{D}$ is defined as:

$$\text{Subset Acc} = \frac{1}{|S|} \sum_{i \in S} \mathbf{1}[\hat{y}_i = y_i]. \quad (3)$$

Table 1 Comparison of multiple evaluation settings on the single character recognition task. Our method consistently outperforms existing baselines across overall, head/mid/tail, and cross-domain (color, rubbing and tracing) evaluations. Bold numbers denote the best results and underline indicates suboptimal results.

	Overall Acc	Balanced Acc	Head Acc	Mid Acc	Tail Acc	Color Acc	Rubbing Acc	Tracing Acc
ABINet [10]	63.64	18.93	71.05	9.90	1.63	50.22	65.98	62.11
PARSeq [9]	60.92	14.32	68.61	3.74	0.19	50.20	62.99	58.28
CLIP-OCR [30]	67.96	25.35	74.80	18.58	5.08	55.14	69.83	69.31
CLIP4STR [11]	76.29	42.38	81.79	40.41	20.68	66.80	77.48	78.86
Ours	78.79	43.23	84.51	41.74	20.31	70.11	79.96	80.43

Full-page Inscription Detection. We evaluate the full-page BI detection performance using the standard Average Precision at a 0.5 IoU threshold (AP_{50}).

Full-page Inscription Recognition. For each page i , we serialize predicted and ground-truth character boxes into sequences $\hat{l}_i$ and l_i using a column-first reading order (columns right-to-left; within-column top-to-bottom), then align $\hat{l}_i$ to l_i via unit-cost Levenshtein to obtain substitution, deletion, and insertion counts (S_i, D_i, I_i) and the reference length $N_i = |l_i|$. Per-page metrics Correct Rate (CR) and Accurate Rate (AR) are defined as:

$$CR_i = \frac{N_i - S_i - D_i}{N_i} \quad (4)$$

$$AR_i = 1 - \frac{S_i + D_i + I_i}{N_i}. \quad (5)$$

For a dataset with M pages, we report macro and micro variants:

$$\text{Macro-CR} = \frac{1}{M} \sum_{i=1}^M CR_i, \text{Macro-AR} = \frac{1}{M} \sum_{i=1}^M AR_i, \quad (6)$$

$$\text{Micro-CR} = 1 - \frac{\sum_{i=1}^M (S_i + D_i)}{\sum_{i=1}^M N_i}, \text{Micro-AR} = 1 - \frac{\sum_{i=1}^M (S_i + D_i + I_i)}{\sum_{i=1}^M N_i}. \quad (7)$$

5.2 Single-Character Recognition

We compare our method with several representative scene text recognition approaches, as summarized in Table 1. Our model achieves the best results on seven of the eight reported metrics, including an Overall Accuracy of **78.79%** and a Balanced Accuracy of **43.23%**, surpassing the previous best (CLIP4STR) by% 2.5 and 0.85%, respectively. For the long-tail evaluation, it reaches **84.51%** on head classes and **41.74%** on mid classes, and remains highly competitive on tail classes with **20.31%**, ranking first in the former two and second in the latter. Across imaging domains, our method consistently delivers superior accuracy with **70.11%** on color images, **79.96%** on rubbings, and **80.43%** on tracings. These results highlight the strong robustness of our approach under class imbalance and diverse visual domains, establishing state-of-the-art performance for single-inscription recognition.

Figure 4 presents correctly recognized character samples across head, mid, and tail frequency groups under diverse imaging conditions. The examples show that our



Fig. 4 Correct recognition examples across frequency groups and domains. The recognition results highlight our model’s robustness to both distribution shifts and domain variations.

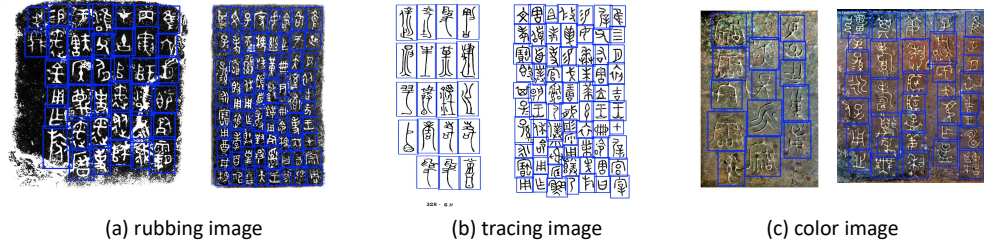


Fig. 5 Detection results of YOLO-v12 on three types of inscription images: (a) rubbing images, (b) tracing images, and (c) color images. The blue bounding boxes denote the detected inscription regions.

model accurately recognizes common characters as well as mid- and low-frequency characters that often appear with severe corrosion, low contrast, or complex textures. Notably, even tail-class samples—where training data are extremely limited and visual patterns are highly degraded—are correctly identified, underscoring the model’s strong generalization ability to rare categories and challenging acquisition domains.

5.3 Full-page Detection and Recognition

We develop a complete full-page bronze inscription (BI) pipeline that first detects inscriptions and then performs end-to-end recognition. For detection, the YOLO-v12 model achieves an AP_{50} of 0.8987, demonstrating strong capability in localizing BI instances despite complex backgrounds and diverse imaging domains. As shown in

Table 2 Performance of full-page bronze inscription detection and recognition. We report the detection metric AP_{50} along with recognition metrics Macro/Micro AR and CR. Bold numbers denote the best results.

Detection Method	AP_{50}	Recognition Method	Macro-AR	Macro-CR	Micro-AR	Micro-CR
YOLO-v12 [40]	0.8987	ABINet [10]	43.45	65.69	54.05	67.32
		PARSeq [9]	40.26	62.54	49.69	62.96
		CLIP-OCR [30]	46.47	68.80	57.23	70.61
		CLIP4STR [11]	48.25	70.63	59.15	72.59
		Ours	49.67	72.05	60.10	73.51

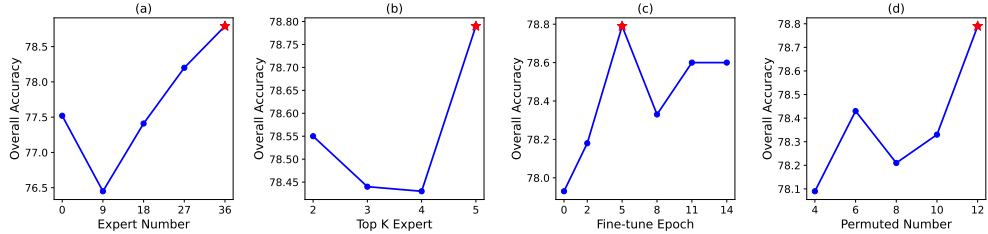


Fig. 6 Overall accuracy versus (a) Varying the number of experts per MoE-Adapter, (b) top-k expert routing, (c) OSF epochs, and (d) PLM permutation count. Red stars denote the optimal settings (36 experts, top-5 routing, 5 OSF epochs, 12 permutations) achieving 78.8% overall accuracy. Expert number, top-k routing, and permutation count show a clear upward correlation with performance.

Figure 5, the model accurately highlights each inscription with bounding boxes across varied modalities and challenging textures.

Building on this detector, we integrate YOLO-v12 with multiple scene text recognition networks to construct the full-page BI recognition pipeline. Table 2 reports the best performance of our pipeline, reaching 49.67% Macro-AR, 72.05% Macro-CR, 60.10% Micro-AR, and 73.51% Micro-CR. These results confirm that the recognition module not only achieves high single-character accuracy but also scales effectively to the full-page setting, validating the robustness of the overall detection–recognition framework.

5.4 Ablation Studies

We perform a series of ablation studies to quantify the contribution of each key component in our framework by overall accuracy metric, as shown in Figure 6. (a) *Number of Experts in MoE Adapters*. Disabling the MoE module (0 experts) yields an overall accuracy of 77.5%. Accuracy dips near 76.5% at 9 experts, then rises steadily to the best performance of 78.8% with 36 experts. This monotonic upward trend after 9 experts indicates that enlarging the expert pool provides richer specialization and stronger representation learning. (b) *Top-k Selection*. As the router’s top-k selection increases from $k = 2$ to $k = 5$, accuracy remains around 78.5% for $k = 2-4$ but reaches 78.8% at $k = 5$. The upward tendency suggests that allowing the router to activate a broader subset of experts facilitates more comprehensive feature aggregation. (c) *Ordered Sequence Fine-tuning (OSF) Epochs*. Without OSF fine-tuning (0 epochs), the

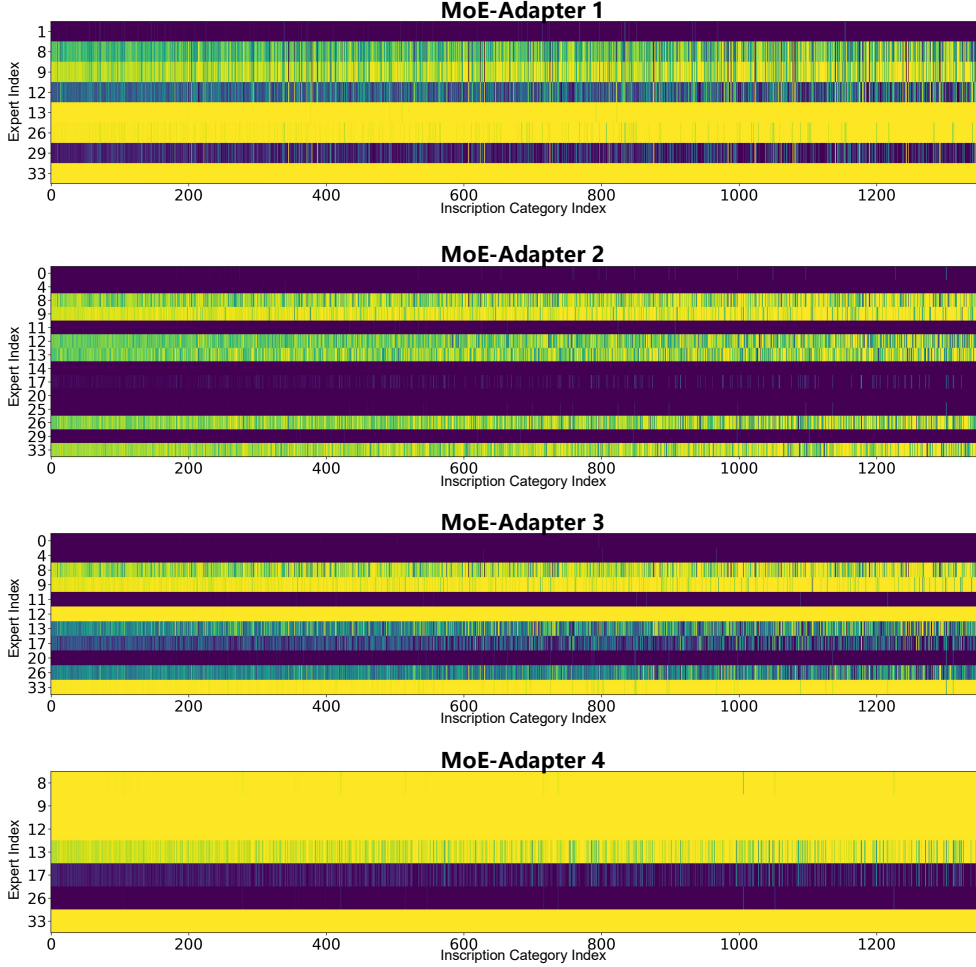


Fig. 7 Visualization of expert activation frequencies across four MoE-Adapters in our Ladder Side framework. The horizontal axis represents character category indices, while the vertical axis denotes expert indices. Bright regions indicate higher activation frequency.

model attains only 77.93% accuracy. Performance climbs with more OSF epochs, peaking near 78.8% at 5 epochs, then shows mild oscillations with further training. This confirms that moderate OSF training effectively reinforces correct character order, while excessive fine-tuning brings no additional gain. (d) *Permuted Sequence Number in PLM*. Increasing the number of permuted sequences from 4 to 12 improves accuracy from 78.1% to 78.8%, highlighting that richer permutation diversity strengthens sequence modeling.

Notably, the relationships in ablation studies (a), (b), and (d) all exhibit a generally ascending trend between parameter magnitude and performance. Although we

observe consistent gains at the tested upper bounds (36 experts, top-5 routing, 12 permutations), resource constraints prevented exploration beyond these settings, leaving open the possibility of further improvements with larger configurations.

5.5 Analysis of Experts Selection

Figure 7 shows the expert activation frequencies of four MoE-Adapters on the test set during inference. Within each adapter, the distribution of activated experts is highly non-uniform: only a small subset of experts are frequently selected, while the majority remain rarely utilized.

When comparing across different adapters, one can observe both overlap and divergence. Certain expert indices (e.g., 9 and 33) are frequently selected in multiple MoE adapters, suggesting that these experts capture universally useful features across character categories. At the same time, different MoE adapters also activate some unique experts internally, indicating that they specialize in complementary subspaces. This inter-adapter diversity suggests that while individual adapters are prone to expert sparsity, the ensemble of multiple adapters ensures broader coverage of the expert pool, thereby enhancing the model’s representation capacity.

6 Conclusion

We presented a large-scale bronze inscription (BI) dataset and a two-stage detection–recognition pipeline that first localizes inscriptions and then transcribes individual characters. To address the key challenges of cross-domain variability, visual degradation, and extreme class imbalance in BI recognition, we propose LadderMoE, a parameter-efficient recognizer that augments a pretrained CLIP encoder with ladder-style mixture-of-experts adapters for dynamic expert specialization. Comprehensive experiments on single-character and full-page tasks confirm that the integrated system consistently surpasses leading scene-text recognition baselines across head, mid, and tail categories and across color, rubbing, and tracing domains, offering a robust and scalable foundation for automatic bronze-inscription recognition and for downstream archaeological analyses.

References

- [1] Guo, R.: A research on an intelligent recognition tool for bronze inscriptions of the shang and zhou dynasties. *Journal of Chinese Writing Systems* 4(4), 271–279 (2020)
- [2] Egorov, A., Egorova, M., Orlova, T.: The use of a comparative analysis of the connection between ancient and modern chinese languages in the process of teaching students chinese characters. In: *2nd International Conference on Education: Current Issues and Digital Technologies (ICECIDT 2022)*, pp. 10–19 (2022). Atlantis Press

- [3] Wolfgang, B.: The language of the bronze inscriptions. Imprints of kinship: Studies of recently discovered bronze inscriptions from ancient China (17), 9 (2017)
- [4] Liu, Z., Wang, X., Yang, C., Liu, J., Yao, X., Xu, Z., Guan, Y.: Oracle character detection based on improved faster r-cnn. In: 2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), pp. 697–700 (2021). IEEE
- [5] Fu, X., Zhou, R., Yang, X., Li, C.: Detecting oracle bone inscriptions via pseudo-category labels. *Heritage Science* **12**(1) (2024)
- [6] Tao, Y., Fu, X., Pang, H., Yang, X., Li, C.: Clustering-based feature representation learning for oracle bone inscriptions detection. *npj Heritage Science* **13**(1), 296 (2025)
- [7] Zheng, Y., Chen, Y., Wang, X., Qi, D., Yan, Y.: Ancient chinese character recognition with improved swin-transformer and flexible data enhancement strategies. *Sensors* **24**(7), 2182 (2024)
- [8] Wu, X., Wang, Z., Ren, P.: Cnn-based bronze inscriptions character recognition. In: 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), pp. 514–519 (2022). IEEE
- [9] Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: European Conference on Computer Vision, pp. 178–196 (2022). Springer
- [10] Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7098–7107 (2021)
- [11] Zhao, S., Quan, R., Zhu, L., Yang, Y.: Clip4str: A simple baseline for scene text recognition with pre-trained vision-language model. *IEEE transactions on Image Processing* (2024)
- [12] Meng, L.: Recognition of oracle bone inscriptions by extracting line features on image processing. In: ICPRAM, pp. 606–611 (2017)
- [13] Wen-Ying, L.I., Bin, C., Chun-Shui, C., Yong-Zhen, H., History, S.O.: A deep learning based method for bronze inscription recognition. *Acta Automatica Sinica* (2018)
- [14] Zhang, Y.-K., Zhang, H., Liu, Y.-G., Yang, Q., Liu, C.-L.: Oracle character recognition by nearest neighbor classification with deep metric learning. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 309–314 (2019). IEEE

- [15] Xia, G., Shang, Z.: Bronze inscription recognition method based on automatic pruning strategy. *Laser & Optoelectronics Progress* **57**(16), 257–264 (2020)
- [16] Atienza, R.: Vision transformer for fast and efficient scene text recognition. In: *International Conference on Document Analysis and Recognition*, pp. 319–334 (2021). Springer
- [17] Aberdam, A., Litman, R., Tsiper, S., Anshel, O., Slossberg, R., Mazor, S., Manmatha, R., Perona, P.: Sequence-to-sequence contrastive learning for text recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15302–15312 (2021)
- [18] Luo, C., Jin, L., Chen, J.: Siman: Exploring self-supervised representation learning of scene text via similarity-aware normalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1039–1048 (2022)
- [19] Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? dataset and model analysis. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019). <https://doi.org/10.1109/iccv.2019.00481> . <http://dx.doi.org/10.1109/iccv.2019.00481>
- [20] Baek, J., Matsui, Y., Aizawa, K.: What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). <https://doi.org/10.1109/cvpr46437.2021.00313> . <http://dx.doi.org/10.1109/cvpr46437.2021.00313>
- [21] Bhunia, A.K., Chowdhury, P.N., Sain, A., Song, Y.-Z.: Towards the unseen: Iterative text recognition by distilling from errors. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14950–14959 (2021)
- [22] Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376 (2006)
- [23] He, P., Huang, W., Qiao, Y., Loy, C., Tang, X.: Reading scene text in deep convolutional sequences. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30 (2016)
- [24] Shi, B., Bai, X., Yao, C.: 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**(11), 2298–2304 (2016)

- [25] Borisjuk, F., Gordo, A., Sivakumar, V.: Rosetta: Large scale system for text detection and recognition in images. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 71–79 (2018)
- [26] Liao, M., Zhang, J., Wan, Z., Xie, F., Liang, J., Lyu, P., Yao, C., Bai, X.: Scene text recognition from two-dimensional perspective. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8714–8721 (2019)
- [27] Wan, Z., He, M., Chen, H., Bai, X., Yao, C.: Textscanner: Reading characters in order for robust scene text recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12120–12127 (2020)
- [28] Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.: Focusing attention: Towards accurate text recognition in natural images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5076–5084 (2017)
- [29] Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence* **41**(9), 2035–2048 (2018)
- [30] Wang, Z., Xie, H., Wang, Y., Xu, J., Zhang, B., Zhang, Y.: Symmetrical linguistic feature distillation with clip for scene text recognition. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 509–518 (2023)
- [31] Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S., Yang, F.: Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review* **58**(8), 227 (2025)
- [32] Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., Zhao, T.: Adaptive budget allocation for parameter-efficient fine-tuning. In: 11th International Conference on Learning Representations, ICLR 2023 (2023)
- [33] Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4582–4597 (2021)
- [34] Hu, E.J., shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022). <https://openreview.net/forum?id=nZeVKeeFYf9>
- [35] Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3045–3059 (2021)

- [36] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. *Neural computation* **3**(1), 79–87 (1991)
- [37] Shi, C., Yang, C., Zhu, X., Wang, J., Wu, T., Li, S., Cai, D., Yang, Y., Meng, Y.: Unchosen experts can contribute too: Unleashing moe models’ power by self-contrast. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*, vol. 37, pp. 136897–136921. Curran Associates, Inc., ??? (2024)
- [38] Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., *et al.*: Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems* **35**, 7103–7114 (2022)
- [39] Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., *et al.*: Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024)
- [40] Tian, Y., Ye, Q., Doermann, D.: Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524* (2025)