

ROOTS: Recognizing Oracle Bone Inscriptions via an Organized Tree Structure

Ziqi Li

Nanjing University of Science and Technology

Zijian Chen

Shanghai Jiao Tong University

Ziyi Yang

Shanghai Jiao Tong University

Zhiji Liu

East China Normal University

Guangtao Zhai

Shanghai Jiao Tong University

Tingzhu Chen

ttingzhuchen@sjtu.edu.cn

Shanghai Jiao Tong University

Article

Keywords: Oracle bone character recognition, Long-tailed distribution, Hierarchical classification, Vision-language model, Superclass generation

Posted Date: May 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9733608/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at npj Heritage Science on September 23rd, 2026. See the published version at <https://doi.org/10.1038/s40494-026-02989-4>.

ROOTS: Recognizing Oracle Bone Inscriptions via an Organized Tree Structure

Ziqi Li^{2†}, Zijian Chen^{3†}, Ziyi Yang¹, Zhiji Liu⁴, Guangtao Zhai³,
Tingzhu Chen^{1*}

¹School of Humanities, Shanghai Jiao Tong University, Shanghai, 200240, China.

²School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China.

³Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai, 200240, China.

⁴Center for the Study and Application of Chinese Characters, East China Normal University, Shanghai, 200241, China.

*Corresponding author(s). E-mail(s): tingzhuchen@sjtu.edu.cn;
Contributing authors: lzq0915@njust.edu.cn; zijian.chen@sjtu.edu.cn;
cestbonyang0422@163.com; 251755227@qq.com;
zhaiguangtao@sjtu.edu.cn;

[†]These authors contributed equally to this work.

Abstract

Oracle bone inscriptions represent the earliest known mature Chinese writing system and constitute invaluable primary sources for studying Shang-dynasty history; yet automated recognition remains hindered by an intrinsically long-tailed class distribution, in which common characters vastly outnumber rare ones, as well as by high inter-class visual similarity. Existing approaches address this imbalance primarily through data augmentation or re-weighting, leaving the underlying label-space skewness largely intact. Here we propose ROOTS (Recognizing Oracle Bone Inscriptions via an Organized Tree Structure), a hierarchical framework that restructures the original 1,771 fine-grained classes into 80 size-balanced superclasses via DINOv2-based prototype extraction and size-aware hierarchical clustering, reducing the mean-to-median class-size ratio from $6.6\times$ to $1.2\times$. Recognition then proceeds coarse-to-fine: a shared root branches into semantic superclasses before resolving individual characters, with a winner-take-all mask confining predictions to the relevant branch. We further integrate

a vision-language pathway in which Qwen3-VL, guided by structured glyph descriptions, performs fine-grained discrimination under stochastic prompt regularization. On a mixed benchmark of 395,845 images spanning five public corpora, ROOTS achieves 95.28% Top-1 accuracy and 99.38% Macro Top-1 accuracy, an absolute gain of 14.2 percentage points over the strongest visual baseline, while lifting tail-class accuracy to 99.83%. These results establish ROOTS as a robust, interpretable solution for long-tailed ancient-script recognition with potential extensibility to other logographic writing systems.

Keywords: Oracle bone character recognition, Long-tailed distribution, Hierarchical classification, Vision-language model, Superclass generation

Introduction

Oracle bone inscriptions, the earliest known mature writing system in China, are not only the origin of Chinese characters but also valuable first-hand historical materials for investigating the politics, economy, culture, and natural environment of the Shang and Zhou periods. Since their discovery at Yinxu in 1899, their decipherment and study have continued for more than a century and have become one of the central research fields in archaeology and paleography. However, among the approximately 4,500 excavated single-character categories, only more than 1,600 have been generally accepted as deciphered by the scholarly community, while nearly two-thirds remain undeciphered. These unknown characters are typically characterized by complex structures, scarce samples, and limited contextual occurrences, making them one of the major bottlenecks in advancing oracle bone inscription research. In this study, we focus on oracle bone character recognition, i.e., fine-grained classification of single-glyph oracle-bone images.

With the rapid development of artificial intelligence, the deep integration of computer vision and natural language processing has opened a new technical route for automated recognition and assisted interpretation of oracle bone inscriptions. Early AI-powered recognition systems established strong image-based pipelines [1]. More recent diffusion-based approaches further advanced visual decipherment [2]. Multi-modal approaches further improved cross-modal understanding [3]. Related work has also examined the perceptual sensitivity of large multimodal models, providing useful context for evaluating fine-grained visual discrimination in multimodal recognition systems [4]. Over the course of this development, oracle bone recognition has evolved from recognizing clean hand-copied tracings produced by later scholars to recognizing noisy characters directly from original rubbings. Recent survey work has further summarized the broader progress of information processing for oracle bone inscriptions from the perspectives of recognition, decipherment, and multimodal understanding [5]. At the same time, more in-depth studies have focused on data imbalance, among which the long-tailed distribution is the most representative challenge.

A long-tailed distribution refers to the phenomenon in which a small number of categories (head classes) contain abundant samples, whereas the vast majority of categories (tail classes) contain only a few samples [6]. In oracle bone studies, some common characters, such as those denoting *person*, *sun*, and *moon*, occur frequently in excavated materials and therefore have sufficient samples. By contrast, characters denoting specific sacrificial rituals, rare species, or geographic names may appear only once or twice among tens of thousands of oracle bones. When such an extremely imbalanced distribution is used to train computer vision recognition models, the resulting models are prone to predictive bias and tend to misclassify tail classes as head classes; this issue has motivated class-balanced loss [7], decoupled training [8], and logit adjustment [9].

To address this problem, many researchers have proposed effective solutions in recent years. One major direction is data augmentation for oracle bone images. CycleGAN-style transfer has been used to map clean handwritten forms to realistic oracle-bone textures [10]. Two-stage GAN pipelines further improve pseudo-sample quality and diversity [11]. In parallel, dynamic augmentation strategies [12] and adversarial augmentation strategies [13] have been introduced to improve robustness under class imbalance. Another line of work focuses on diffusion models and structure-aware sample synthesis. Compared with GANs, diffusion models generally provide better image diversity and fidelity for delicate oracle bone structures. Representative approaches include structure-aware generation/reconstruction pipelines [14], which effectively assist downstream recognition.

The overall ROOTS framework is illustrated in Fig. 1.

The methods discussed above mainly address the long-tailed problem in oracle bone character recognition through data augmentation. On the algorithmic side, many researchers have also proposed effective solutions, including AI-powered recognition pipelines [1], imbalance-aware training strategies and benchmarks [15, 16], and domain-transfer-based recognition frameworks [17]. These methods improve recognition under sparse supervision, but the core label-space imbalance remains an important bottleneck. To our knowledge, existing studies mainly mitigate long-tailed effects through data augmentation or imbalance-aware learning, rather than restructuring the original label space itself, so the recognition difficulty induced by the intrinsic long-tailed distribution is not directly reduced at the label-space level. One promising direction is therefore to replace flat recognition with a superclass-based hierarchical formulation. Rather than crudely collapsing categories, hierarchical classification reorganizes the original flat label space into a coarse-to-fine structure, so that prediction proceeds from broader groups to finer character identities [18]. For oracle bone recognition, such a reformulation is particularly attractive because it narrows the candidate space at each decision stage, weakens the direct competition between heavily sampled head classes and severely underrepresented tail classes, and separates globally distinct groups before discriminating among highly similar glyphs within each group. In addition, the intermediate superclass prediction provides an explicit decision path for the final output, making the recognition process more interpretable than a single flat classifier.

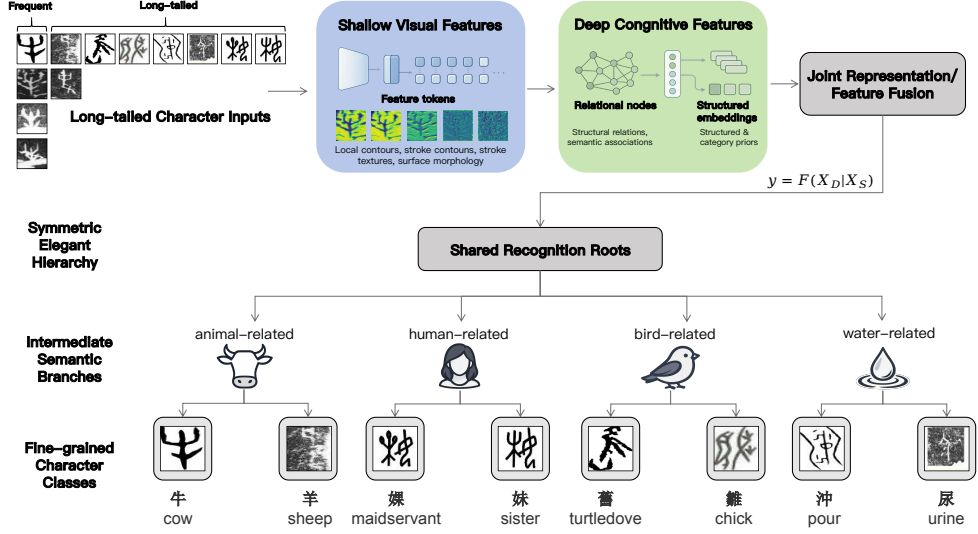


Fig. 1 Overview of the ROOTS framework for hierarchical oracle bone character recognition. The upper section illustrates the feature extraction pipeline: oracle bone images with a long-tailed class distribution are processed through shallow visual features (local contours, stroke contours, and surface morphology) and deep cognitive features (relational nodes and structured embeddings), which are then fused into a joint representation $y = F(X_D | X_S)$. The lower section depicts the coarse-to-fine classification hierarchy, where a shared recognition root branches into intermediate semantic superclasses (e.g., animal-, human-, bird-, and water-related forms), each further subdivided into fine-grained character classes.

Hierarchical strategies have shown clear value across recognition problems that share the structural difficulties of oracle bone character classification. In large-scale image classification, reorganizing categories into a learned structure improves both efficiency and predictive performance by reducing the effective inference space [19]. In fine-grained and taxonomy-aware recognition such as plant species identification, coarse grouping helps reduce confusion among subtly different categories [20]. In medical image diagnosis, parent-child label organization supports fine-grained subtype discrimination and more interpretable decisions [21]. In document or script recognition, multi-level semantic organization improves robustness under substantial visual variation and inter-class similarity [22]. These applications all involve large label spaces, strong fine-grained resemblance, or category imbalance, which closely parallels the challenges of oracle bone recognition. Motivated by these successes, we investigate whether oracle bone character recognition can also benefit from a learned superclass hierarchy and, based on this intuition, reorganize the original label space and perform recognition in a coarse-to-fine manner.

To address these limitations, we propose the ROOTS framework, which first generates size-balanced superclasses through a pipeline of feature extraction, representative-character selection, and hierarchical clustering, and then performs coarse-to-fine

recognition through both a purely visual implementation and a multimodal implementation based on large vision-language models, both of which mimic the human tree-structured cognitive process. Under the newly constructed superclass system, the data distribution becomes substantially more balanced, enabling end-to-end oracle bone image recognition. The main contributions of this study are as follows:

1. We reorganize 1,771 fine-grained oracle bone character classes into 80 size-balanced superclasses using DINOv2 prototypes and size-aware hierarchical clustering, thereby reducing structural label-space imbalance.
2. We design a coarse-to-fine ROOTS framework with hierarchical prediction heads and winner-take-all masking, and we instantiate it in both visual-only and VLM-based settings.
3. On the mixed dataset of 395,845 images, ROOTS improves Macro Top-1 accuracy from 85.16% achieved by the pre-trained ConvNeXt-Base baseline to 99.38% achieved by ROOTS with Qwen3-VL-2B, with substantial gains on tail classes.

Methods

Task Definition

In this section, we formally define the problem of long-tailed oracle bone character recognition and introduce our hierarchical task formulation.

Standard Recognition Task. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote the oracle bone image dataset, where $x_i \in \mathcal{X}$ is the input image and $y_i \in \mathcal{Y} = \{1, 2, \dots, K\}$ is the corresponding fine-grained class label. In our mixed dataset, $K = 1771$. Let N_j denote the number of training samples belonging to class $j \in \mathcal{Y}$. Since oracle bone data are intrinsically long-tailed [6], these class frequencies satisfy $N_1 \geq N_2 \geq \dots \geq N_K$ after sorting, with a substantial gap between frequent head classes and rare tail classes. The standard objective is to learn a flat mapping $F : \mathcal{X} \rightarrow \mathcal{Y}$; however, directly optimizing on such an imbalanced fine-grained label space makes the predictor highly vulnerable to head-class bias [7].

Superclass Formulation. To mitigate this structural imbalance, we reorganize the original fine-grained label space into a superclass space $\mathcal{Y}_c = \{1, 2, \dots, K_c\}$, where $K_c \ll K$. In this study, $K_c = 80$. We define a surjective mapping $\phi : \mathcal{Y} \rightarrow \mathcal{Y}_c$ that assigns each fine-grained class to a semantically related and size-balanced superclass. Accordingly, for each superclass $k \in \mathcal{Y}_c$, we denote by $\mathcal{C}(k) = \{y \in \mathcal{Y} \mid \phi(y) = k\}$ the subset of fine-grained classes contained in that branch. This reformulation converts the original flat recognition problem into a structured prediction problem over a coarse-to-fine hierarchy.

Hierarchical Recognition Task. Based on the generated superclass system, we redefine recognition as two sequential sub-tasks: coarse-level prediction over $\mathcal{Y}_c$ and fine-grained prediction conditioned on the selected superclass. The model first estimates the superclass posterior $P(y_c = k \mid x)$, where y_c denotes the superclass label, and determines the optimal superclass

$$\hat{y}_c = \arg \max_{k \in \mathcal{Y}_c} P(y_c = k \mid x), \quad (1)$$

where $\hat{y}_c$ denotes the predicted superclass index. The final prediction is then restricted to the candidate subset covered by $\hat{y}_c$:

$$\hat{y} = \arg \max_{y \in \mathcal{C}(\hat{y}_c)} P(y \mid x, \hat{y}_c). \quad (2)$$

Here, $\hat{y}$ denotes the final predicted fine-grained class. This coarse-to-fine formulation explicitly narrows the candidate space at each decision stage and weakens the direct structural competition between heavily sampled head classes and severely underrepresented tail classes.

Superclass Generation

To handle the extreme long-tailed distribution and high visual similarity among oracle-bone glyphs, we propose a size-aware hierarchical clustering algorithm. The purpose of this process is to aggregate fine-grained classes into semantically related and size-balanced superclasses, thereby reducing the complexity of downstream recognition. This design is consistent with prior findings on hierarchical classification in large taxonomies [23] and hierarchy-aware category structure learning [19].

The superclass generation workflow is shown in Fig. 2.

First, we use the pre-trained visual DINOv2 model to extract deep features for each class. Given the original dataset, we adopt pre-trained DINOv2 (ViT-L/14) as the feature extractor [24]. All images are first resized to 518×518 pixels by bicubic interpolation and then normalized. For each class c , its feature prototype μ_c is obtained by averaging the feature vectors $f(x_i)$ of all images in that class:

$$\mu_c = \frac{1}{N_c} \sum_{i=1}^{N_c} f(x_i) \quad (3)$$

where $f(\cdot)$ denotes the pre-trained DINOv2 feature extractor, μ_c is the prototype of class c , and N_c denotes the number of image samples contained in class c . In addition, we select the image whose Euclidean distance to the prototype μ_c is the smallest as the representative sample of that class. A prototype matrix is then constructed for subsequent clustering.

Traditional Ward clustering tends to minimize the within-cluster sum-of-squares error [25]. However, when applied to highly imbalanced data, it can easily produce extremely large “giant clusters,” which is unfavorable for downstream classifier training. To handle this issue, we introduce a size penalty factor α and redefine the merging cost function.

In the superclass generation procedure, the merge cost function for two clusters C_i and C_j is defined as

$$J(C_i, C_j) = \Delta E_{\text{Ward}}(C_i, C_j) \times (|C_i| + |C_j|)^\alpha \quad (4)$$

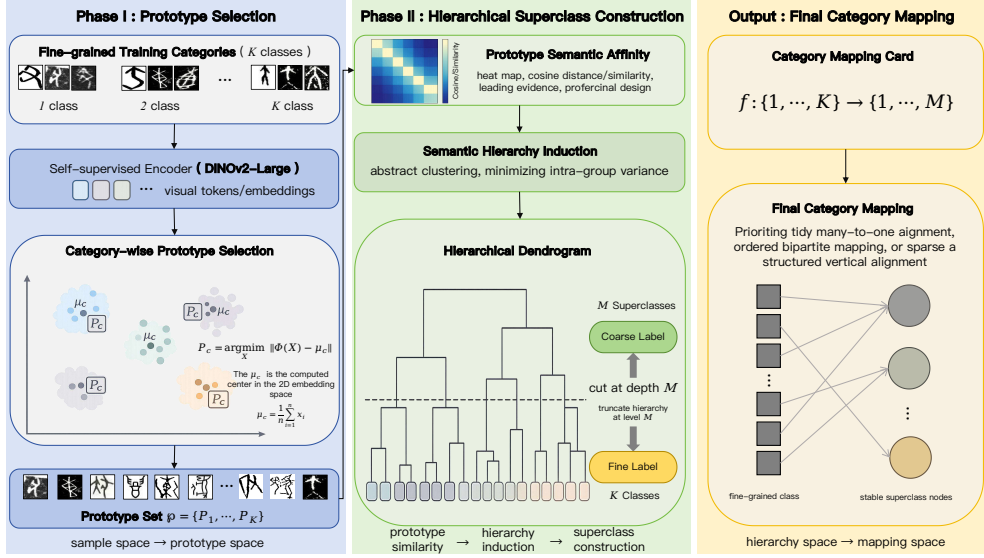


Fig. 2 Workflow of superclass generation. The pipeline consists of three stages. **Phase I (Prototype Selection)**: Fine-grained character images are encoded via a self-supervised DINOv2-Large encoder, and the sample nearest to each class centroid μ_c in the embedding space is selected as the prototype P_c , yielding a prototype set $\varphi = \{P_1, \dots, P_K\}$. **Phase II (Hierarchical Superclass Construction)**: A semantic affinity matrix is computed using cosine similarity among prototypes, followed by hierarchical clustering to produce a dendrogram; cutting the dendrogram to obtain M clusters partitions the K fine-grained classes into M size-balanced superclasses. **Output (Final Category Mapping)**: A many-to-one mapping $f: \{1, \dots, K\} \rightarrow \{1, \dots, M\}$ assigns each fine-grained class to its corresponding superclass, transforming the original long-tailed label space into a more balanced hierarchical structure.

where ΔE_{Ward} is the standard Ward distance:

$$\Delta E_{\text{Ward}}(C_i, C_j) = \frac{|C_i| \cdot |C_j|}{|C_i| + |C_j|} \|\mu_i - \mu_j\|^2 \quad (5)$$

Here, $J(C_i, C_j)$ denotes the penalized cost of merging clusters C_i and C_j , $|C_i|$ and $|C_j|$ denote the numbers of image samples contained in the two clusters, and μ_i and μ_j are their corresponding cluster prototypes. By introducing the penalty term $(|C_i| + |C_j|)^\alpha$, the algorithm dynamically increases the cost of merging clusters that already contain many samples. When $\alpha > 0$, the continued expansion of large clusters is suppressed, forcing smaller clusters to merge first and thereby automatically balancing the sample sizes of the resulting superclasses during clustering.

Purely Visual Implementation of ROOTS

Building upon the established hierarchical task definition, we first instantiate the ROOTS framework using a purely visual architecture. This implementation aims

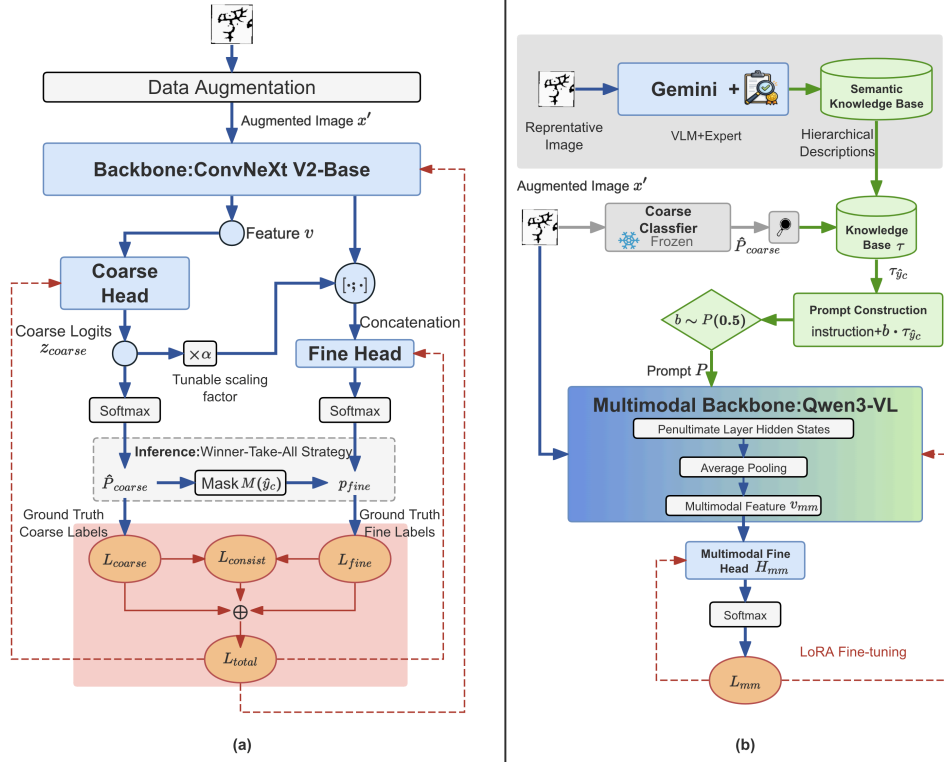


Fig. 3 Two implementation pipelines of the ROOTS framework. **(a) Purely visual pipeline:** An augmented image x' is passed through a ConvNeXt V2-Base backbone to extract feature v . A coarse head produces superclass logits z_{coarse} , which are scaled by a tunable factor α and concatenated with v to feed the fine head. During inference, a winner-take-all (WTA) mask $M(\hat{y}_c)$ constrains fine-grained predictions to the predicted superclass branch. The model is trained end-to-end with a joint loss $L_{\text{total}} = L_{\text{coarse}} + L_{\text{consist}} + L_{\text{fine}}$. **(b) Multimodal pipeline:** A semantic knowledge base is constructed offline using Gemini with expert refinement. During inference, a frozen coarse classifier predicts $\hat{P}_{\text{coarse}}$, which retrieves superclass-specific descriptions $\tau_{\hat{y}_c}$ from the knowledge base. A stochastic prompt $P = \text{instruction} + b \cdot \tau_{\hat{y}_c}$ (where $b \sim \text{Bernoulli}(0.5)$) is constructed to prevent over-reliance on text. The prompt and image are fed into Qwen3-VL, fine-tuned via LoRA, to produce the final prediction.

to validate, using strictly visual evidence, the effectiveness of the superclass structure, independent of external semantic knowledge bases. The overall visual pipeline is designed as an end-to-end learning system comprising three core components. First, a deep convolutional backbone extracts robust visual representations, which are subsequently processed by dual prediction heads for coarse-to-fine hierarchical classification. Second, a winner-take-all (WTA) constraint mechanism is introduced to strictly prune the fine-grained candidate space during inference based on the coarse prediction.

Finally, the network is trained using a joint optimization objective that enforces semantic consistency across different hierarchy levels. The specific formulations of these modules are detailed in the following subsections.

The two ROOTS implementation pipelines are shown in Fig. 3.

Feature extraction and hierarchical prediction heads

Given an input oracle bone image x , an augmented image $x' = T(x)$ is first obtained through random augmentation operators. We use the ConvNeXt family [26], specifically ConvNeXt V2-Base with FCMAE pre-training [27], as the backbone network f_θ . After global average pooling, the visual representation vector $v \in \mathbb{R}^d$ is extracted as follows:

$$v = \text{GAP}(f_\theta(x')) \quad (6)$$

where $\text{GAP}(\cdot)$ denotes global average pooling and d is the feature dimension.

This branch adopts a dual-head architecture for hierarchical classification. The coarse head $\mathcal{H}_c$ first computes the superclass logits z_{coarse} and the corresponding probability distribution P_{coarse} :

$$z_{\text{coarse}} = \mathcal{H}_c(v), \quad P_{\text{coarse}} = \text{softmax}(z_{\text{coarse}}) \quad (7)$$

where $\mathcal{H}_c$ is the coarse prediction head, $z_{\text{coarse}} \in \mathbb{R}^{K_c}$ denotes the superclass logits, $P_{\text{coarse}} \in \Delta^{K_c-1}$ is the corresponding superclass probability vector, K_c denotes the number of superclasses, and Δ^{K_c-1} denotes the $(K_c - 1)$ -dimensional probability simplex.

To enable information transfer across hierarchy levels, the coarse prediction is fed back into the fine-grained feature space. By introducing a tunable scaling factor β , we construct a fused feature:

$$v_{\text{fine}} = [v \parallel \beta \cdot z_{\text{coarse}}] \quad (8)$$

where v_{fine} is the fused feature for fine-grained prediction, β controls the strength of the fed-back coarse signal, and $\parallel$ denotes vector concatenation. The fine head $\mathcal{H}_f$ then outputs the final fine-grained class probability distribution:

$$p_{\text{fine}} = \text{softmax}(\mathcal{H}_f(v_{\text{fine}})) \quad (9)$$

where $\mathcal{H}_f$ is the fine prediction head and $p_{\text{fine}} \in \Delta^{K-1}$ denotes the probability distribution over the K fine-grained classes.

Winner-take-all (WTA) inference strategy

During inference, to effectively narrow the search space, we introduce a winner-take-all (WTA) constraint mechanism, which is consistent with the general motivation of hierarchical decision making in large-scale taxonomies [23]. The coarse head first determines the predicted superclass:

$$\hat{y}_c = \text{argmax}_{k \in \mathcal{Y}_c} P_{\text{coarse},k} \quad (10)$$

where $P_{\text{coarse},k}$ denotes the predicted probability assigned to the k -th superclass.

Based on the ROOTS superclass hierarchy, a mask vector $M(\hat{y}_c) \in \{0, 1\}^K$ is constructed, where K is the number of fine-grained classes. By applying element-wise multiplication (Hadamard product) between this mask and the fine-grained probability distribution, hard pruning of the candidate space is achieved:

$$\hat{y}_{\text{fine}} = \operatorname{argmax}_{i \in \mathcal{Y}} (p_{\text{fine},i} M_i(\hat{y}_c)) \quad (11)$$

where $\hat{y}_{\text{fine}}$ denotes the final fine-grained prediction, $p_{\text{fine},i}$ is the predicted probability of the i -th fine-grained class, and $M_i(\hat{y}_c)$ is the i -th binary entry of the mask induced by superclass $\hat{y}_c$.

This strategy restricts fine-grained inference to the candidate subset covered by the predicted superclass. In practice, it reduces interference from visually similar classes outside that branch.

Joint optimization objective

We optimize the visual branch end-to-end using a weighted combination of three loss functions: (1) the coarse classification loss $\mathcal{L}_{\text{coarse}}$, i.e., the cross-entropy between P_{coarse} and the ground-truth superclass label; (2) the fine classification loss $\mathcal{L}_{\text{fine}}$, i.e., the cross-entropy between p_{fine} and the ground-truth fine-grained label; and (3) the hierarchical consistency loss $\mathcal{L}_{\text{consist}}$, which constrains the coarse-level marginal induced by p_{fine} to match the ground-truth superclass:

$$\tilde{p}_{\text{coarse},k} = \sum_{i \in \mathcal{C}(k)} p_{\text{fine},i}, \quad \mathcal{L}_{\text{consist}} = \text{CE}(\tilde{p}_{\text{coarse}}, y_c^*) \quad (12)$$

where $\tilde{p}_{\text{coarse},k}$ is the coarse marginal probability obtained by summing the fine-grained probabilities within superclass k , $\tilde{p}_{\text{coarse}}$ is the resulting marginal distribution over all superclasses, $\mathcal{C}(k)$ is the set of fine-grained classes belonging to superclass k , y_c^* denotes the ground-truth superclass label, and $\text{CE}(\cdot, \cdot)$ is the cross-entropy loss.

The final loss function is defined as

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{coarse}} + \lambda_2 \mathcal{L}_{\text{fine}} + \lambda_3 \mathcal{L}_{\text{consist}}, \quad \lambda_1, \lambda_2, \lambda_3 \geq 0 \quad (13)$$

where $\mathcal{L}_{\text{total}}$ is the overall training objective and λ_1 , λ_2 , and λ_3 are non-negative weighting coefficients.

Multimodal Implementation of ROOTS

While the purely visual implementation of ROOTS effectively navigates the superclass hierarchy to reduce label-space imbalance, exclusive reliance on visual features can still be limiting for distinguishing extremely rare tail classes that exhibit high inter-class similarity. To further enhance discriminative capacity, we introduce a multimodal implementation of ROOTS. This approach systematically integrates visual evidence with structured semantic descriptions, leveraging the cross-modal reasoning capabilities of large vision-language models to mimic a tree-structured cognitive process.

The multimodal pipeline is structured into two primary components. First, an offline semantic knowledge base is constructed to explicitly encode the glyph-formation rationale and evolutionary characteristics of each fine-grained character. Second, an online inference mechanism is designed using a random prompting strategy; this mechanism retrieves relevant semantic priors based on the coarse-level prediction to guide the fine-grained multimodal backbone. This dual-stage design forces the model to maintain robust visual discrimination while dynamically benefiting from external expert knowledge. The specific methodologies for knowledge base construction and prompt formulation are detailed in the subsequent subsections.

Semantic knowledge base construction

In the offline stage, we use Gemini 2.5-Flash as the description generator G_{desc} to produce initial hierarchical descriptions for each fine-grained class, which are subsequently reviewed and refined by oracle bone specialists to ensure philological accuracy and semantic consistency. For the representative image of class c , the semantic entry τ_c is constructed as

$$\tau_c = G_{\text{desc}}(\text{Image}_{\text{rep},c}, \text{Instruction}_{\text{hier}}) \quad (14)$$

where $\text{Image}_{\text{rep},c}$ denotes the representative image of class c , $\text{Instruction}_{\text{hier}}$ is the hierarchy-aware prompt used for description generation, and τ_c is the resulting semantic description associated with class c .

All entries are stored in the semantic knowledge base K . Each entry covers the glyph-formation rationale and evolutionary characteristics of the corresponding character.

Random prompting mechanism for inference

During online training and inference, a frozen coarse classifier first predicts the super-class $\hat{y}_c$ of the current image. We then retrieve the fine-grained description set under that branch, i.e., $\mathcal{T}_{\hat{y}_c} = \{\tau_i \mid i \in \mathcal{C}(\hat{y}_c)\}$, from the knowledge base K . To prevent the model from over-relying on text, we adopt stochastic prompt regularization with $\tilde{\tau} \sim \text{Uniform}(\mathcal{T}_{\hat{y}_c})$ and $b \sim \text{Bernoulli}(0.5)$. The multimodal prompt P is defined as

$$P = \text{Instruction} \oplus (b \cdot \tilde{\tau}) \quad (15)$$

where Instruction denotes the base instruction template, $\tilde{\tau}$ is the sampled superclass-consistent description, b is the binary gate controlling whether the sampled description is injected, $\oplus$ denotes text concatenation, and P is the final prompt fed to the multimodal model.

When $b = 1$, the input contains rich semantic priors; when $b = 0$, the model receives only the basic instruction. This strategy forces the model to maintain robust visual discrimination even in the absence of external knowledge.

Results

Composition of the original dataset

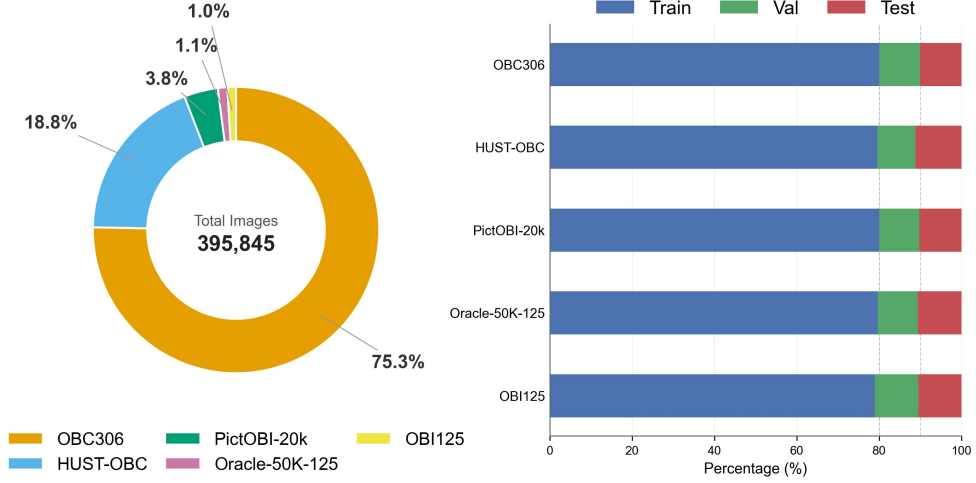


Fig. 4 Statistical overview of dataset composition. **Left:** A donut chart showing the composition of the mixed dataset (395,845 images in total), comprising five public oracle bone corpora: OBC306 (75.3%), HUST-OBC (18.8%), PictOBI-20k (3.8%), Oracle-50K-125 (1.1%), and OBI125 (1.0%). **Right:** Stacked bar chart illustrating the unified 80:10:10 train/validation/test split applied consistently across all subsets to ensure non-overlapping partitions and fair evaluation.

Many studies have shown that training on mixed datasets can enhance model generalization [12]. Accordingly, we adopt the same strategy and merge multiple publicly available oracle bone datasets [28]. The resulting dataset consists of five major subsets and exhibits a clear long-tailed distribution. OBC306 is the core component, accounting for 75.2% of the data [29], followed by HUST-OBC at 18.9% [28]. The remaining sources are PictOBI-20k (3.8%) [30], Oracle-50K-125 (1.1%) [12], and OBI125 (1.0%) [31]. This multi-source fusion covers oracle bone forms from different periods and excavation backgrounds and substantially improves the generalization ability of the dataset. To ensure rigorous and fair model evaluation, we apply a unified split ratio to all five subsets. As shown in Fig. 4, each subset is randomly divided into 80% training, 10% validation, and 10% test data. This consistent partitioning not only stabilizes training but also ensures that the test results accurately reflect model performance across different data sources. Special care is taken in splitting long-tailed classes: when a class contains more than three samples, the train/validation/test subsets are ensured to contain at least one sample each; when a class contains exactly two samples, one is assigned to training and the other to testing; when a class contains only one sample, it is assigned to the training set. Classes with only one sample are excluded from validation/test-based class-wise evaluation metrics.

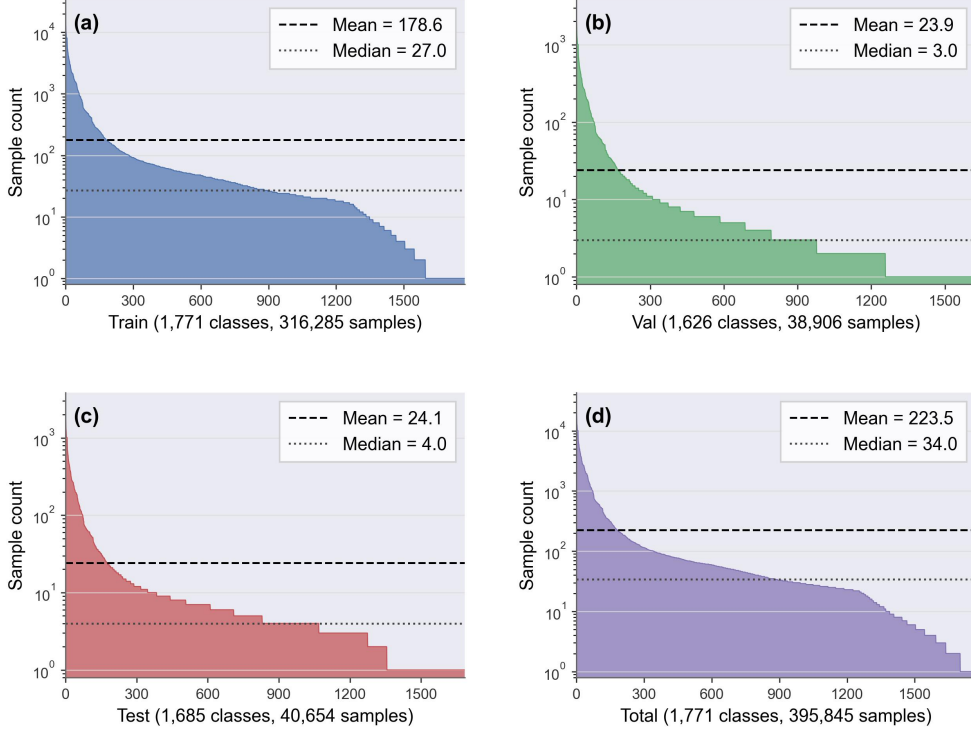


Fig. 5 Long-tailed class distribution of the dataset. Sample counts per class (y -axis, log scale) are plotted in descending order for (a) training set (1,771 classes, 316,285 samples), (b) validation set (1,626 classes, 38,906 samples), (c) test set (1,685 classes, 40,654 samples), and (d) full dataset (1,771 classes, 395,845 samples). Dashed and dotted lines indicate the mean and median, respectively. Across all splits, the mean exceeds the median by approximately 6–8 $\times$ (e.g., 223.5 vs. 34.0 for the full set), confirming a severe long-tailed imbalance where a small number of head classes dominate the sample distribution.

A statistical analysis of 395,845 oracle bone images shows that the dataset exhibits extreme class imbalance and a typical long-tailed distribution. In the complete dataset containing 1,771 classes, the average number of samples per class is 223.5, whereas the median is only 34.0. The fact that the mean is far larger than the median directly reveals that a small number of head classes occupy most of the samples, while a vast number of tail classes contain very few samples. The same phenomenon is also observed in the validation and test sets.

Superclass dataset

Under the original distribution, the dataset is extremely imbalanced. A large statistical gap exists between the overall mean (223.5) and median (34.0). In the training set in particular, the median is only 27.0, indicating that more than half of the classes have extremely few samples. This sparse long-tailed pattern severely hinders classifier learning. After introducing the superclass structure, however, the balance of the

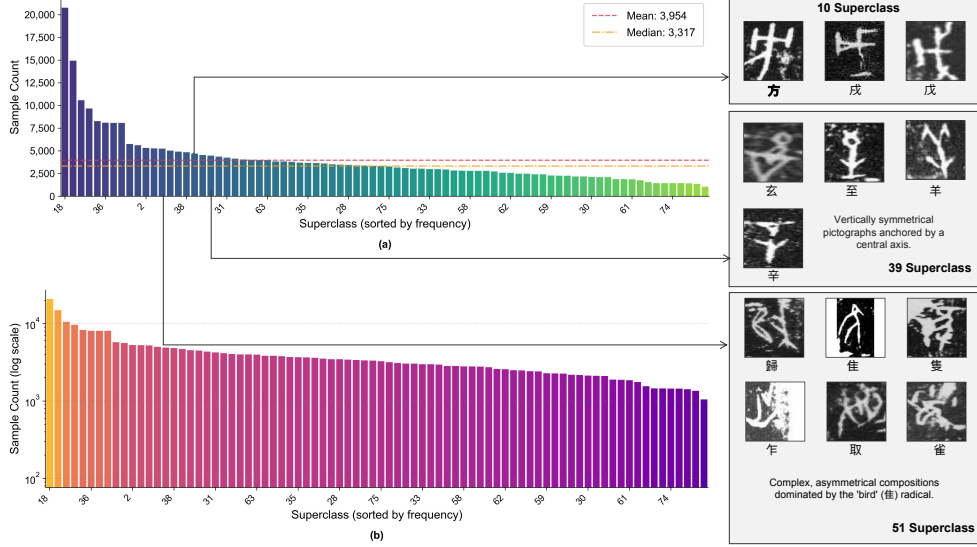


Fig. 6 Distribution statistics of the superclass dataset. (a) Bar chart showing sample counts across 80 superclasses sorted by frequency (linear scale). The mean (3,954) and median (3,317) differ by only 1.19 $\times$, compared to 6–8 $\times$ in the original fine-grained distribution (cf. Fig. 5), indicating substantially improved class balance. (b) The same distribution on a logarithmic scale, where approximately 90% of superclasses fall within the 10^3 – 5×10^3 range, confirming reduced variance. Right panels illustrate three representative superclasses: Superclass 10 (horizontal-bar structures), Superclass 39 (vertically symmetrical pictographs anchored by a central axis), and Superclass 51 (complex compositions with the bird radical). These examples demonstrate that the generated superclasses preserve meaningful morphological and semantic coherence.

dataset is markedly improved. As shown in Fig. 6, the training set (316,285 samples) exhibits a more statistically meaningful stepwise distribution over 80 superclasses. The logarithmic-scale comparison in Fig. 6 (b) further reveals the extent of this balancing effect. In log space, except for a few head superclasses, roughly 90% of the superclasses remain within the same order-of-magnitude interval between 10^3 and 5×10^3 samples. This convergence alleviates classifier bias toward high-frequency classes. Notably, the superclass distribution still retains pronounced head characteristics, such as superclass S18 (20,797 samples). This reflects the true archaeological distribution of oracle bone characters: some core classes occur extremely frequently in excavated corpora and are semantically too independent to be further merged under the learned hierarchy. Rather than forcibly down-sampling such classes, ROOTS uses the hierarchical mask mechanism (WTA) to narrow the subsequent fine-grained search space.

Multimodal Dataset Construction

As shown in Fig. 7, we design a hierarchical multimodal instruction-construction scheme for large vision-language models such as Qwen2-VL and Qwen3-VL [32] under long-tailed distributions. The goal is to guide the model, through a structured knowledge base K , to keep recognition anchored in visual evidence while reducing

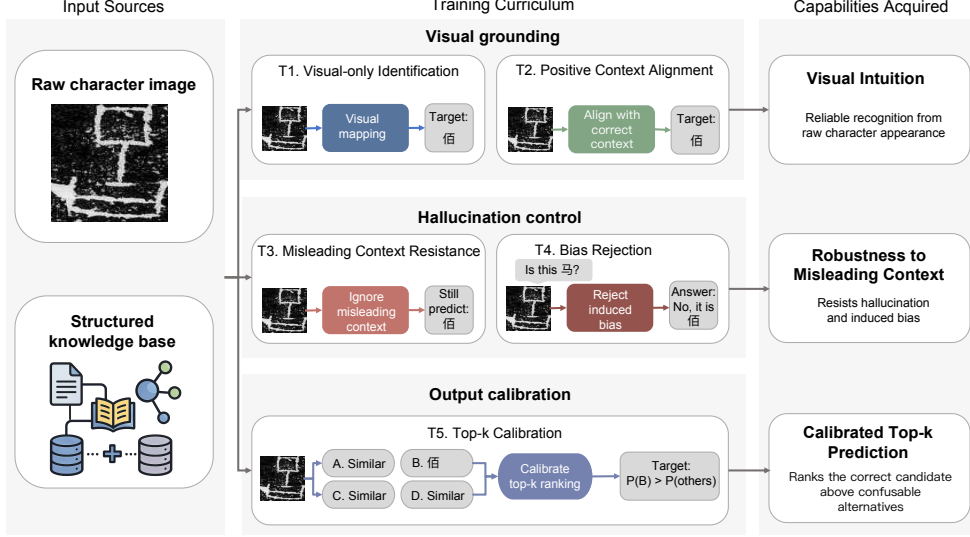


Fig. 7 Pipeline for multimodal training data construction. The framework takes two input sources: raw character images and a structured knowledge base, and defines five training tasks (T1–T5) organized into three capability modules. **Visual grounding:** T1 (Visual-only Identification) establishes a purely visual baseline $y = f(x)$; T2 (Positive Context Alignment) pairs images with correct semantic descriptions to strengthen vision-language alignment. **Hallucination control:** T3 (Misleading Context Resistance) trains the model to ignore incorrect textual cues; T4 (Bias Rejection) teaches the model to resist leading questions (e.g., “Is this 马?”). **Output calibration:** T5 (Top- k Calibration) presents confusable candidates in multiple-choice format to ensure the correct class ranks above visually similar alternatives. Together, these tasks yield three core capabilities: visual intuition, robustness to misleading context, and calibrated top- k prediction.

text-induced bias. This design is consistent with recent oracle-bone multimodal studies [33].

Given the extreme imbalance of the oracle bone dataset (median 34.0 samples per class, while head classes exceed 20,000 samples), we adopt a differentiated task-construction strategy. For the top 20% most frequent classes, we use Task 1 and Task 2 to strengthen alignment between visual features and domain semantics. For the remaining 80% of medium- and tail-frequency classes, we use the full task set (T1–T5) because visual evidence is much sparser. By introducing distracting text and multiple-choice induction, the model is trained to preserve visual consistency even in low-resource settings.

We define five core fine-tuning tasks as follows.

1. **T1: Direct recognition.** Only the original image x is provided, and the model is trained to perform a purely visual mapping $y = f(x)$, thereby establishing a visually grounded baseline.
2. **T2: Positive context injection.** The image x is paired with the correct expert-knowledge entry k^+ , i.e., $y = f(x, k^+)$, to strengthen alignment between visual representations and class semantics.

3. **T3: Negative context injection.** Distracting or incorrect text k^- is introduced, and the model is forced to ignore misleading information while maintaining its visually grounded judgment, i.e., $y = f(x, k^-) = \text{Target}$.
4. **T4: Anti-hallucination robustness training.** Leading question forms, such as “Is this the character for horse?”, are used to train the model to resist misleading prompts and output a calibrated correct result.
5. **T5: Top- k -induced calibration.** Three visually similar candidate characters are provided in multiple-choice form, and the model is required to make a precise discrimination within a semantically confusing candidate space.

Oracle Bone Character Recognition Results

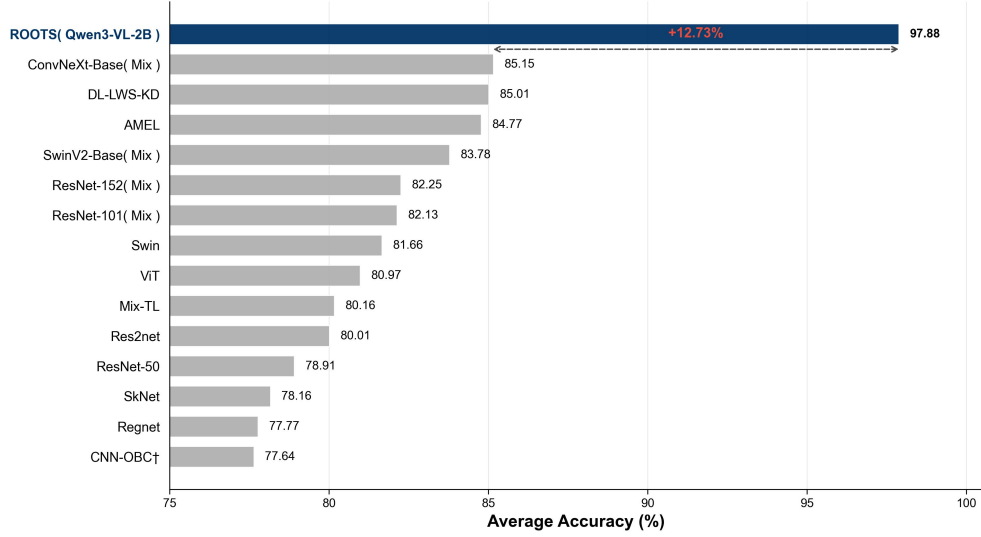


Fig. 8 Comparison of average accuracy on the OBC306 benchmark. The horizontal bar chart ranks 15 methods by average accuracy (x -axis starts at 75% for visual clarity). ROOTS (Qwen3-VL-2B) achieves the highest accuracy of 97.88%, outperforming the second-best method, ConvNeXt-Base (Mix), by +12.73 percentage points. Methods labeled “(Mix)” are trained on the mixed multi-source dataset. The comparison spans CNN-based baselines (ResNet, Res2Net, RegNet, SkNet, CNN-OBC), Transformer-based models (ViT, Swin, SwinV2), long-tail-specific methods (DL-LWS-KD, AMEL, Mix-TL), and the proposed VLM-based ROOTS framework.

Figure 8 reports the average accuracy of mainstream deep learning models on OBC306. This metric provides a direct summary of overall recognition quality under a unified evaluation setting. Methods whose names include “Mix” are trained on the mixed dataset, while the others follow their corresponding standard training settings. The results show a clear stepwise pattern, indicating that architectural and training-strategy upgrades consistently improve long-tailed oracle bone character recognition.

As shown in Fig. 8, ROOTS (Qwen3-VL-2B), a vision-language model based on large-scale pre-training [32], achieves the best overall result, with an average accuracy of 97.88%. This result suggests that structured semantic priors provide additional benefits beyond purely visual feature extraction for this task. Improved convolutional or attention-based methods, including ConvNeXt [26, 27], DL-LWS-KD [34], and AMEL [35], form the second tier. Earlier benchmark models, such as CNN-OBC [29], RegNet [36], and ResNet-50, remain markedly lower. The benchmark setting also includes representative CNN backbones such as SkNet [37] and Res2Net [38]. Overall, the transition from conventional CNNs to Swin Transformer [39] and then to large vision-language models [32] is accompanied by steady performance gains.

Table 1 Overall comparison on the mixed dataset. The **best** and second-best results in each column are highlighted. M-Top-1 denotes Macro Top-1 accuracy.

Method	Top-1 (%)	Top-5 (%)	M-Top-1 (%)
<i>Pure Vision (Scratch)</i>			
ConvNeXt-Base	92.76	98.20	81.54
ResNet-101	92.66	98.23	82.23
ResNet-152	92.58	98.18	82.21
SwinV2-Base	93.85	98.42	84.98
<i>Pure Vision (Pre-trained)</i>			
ConvNeXt-Base	94.12	98.59	85.16
ResNet-101	92.32	98.08	81.16
ResNet-152	92.49	98.16	81.26
SwinV2-Base	93.85	98.42	84.98
ROOTS (ConvNeXt-Base)	94.37	99.23	<u>91.11</u>
<i>VLM-based</i>			
ROOTS (Qwen3-VL-2B)	95.28	98.42	99.38

To better analyze long-tailed recognition on the mixed dataset, we evaluate multiple metrics for each method. As shown in Table 1, ROOTS (Qwen3-VL-2B) achieves a Top-1 accuracy of 95.28% and a Macro Top-1 accuracy of 99.38%. Here, Macro Top-1 is computed over all evaluable classes on the mixed test set.

The clearest gain is observed in Macro Top-1. Compared with the best-performing purely visual model (pre-trained ConvNeXt-Base), ROOTS (Qwen3-VL-2B) improves Macro Top-1 by 14.22 percentage points (99.38% vs. 85.16%). Since Macro Top-1 weights all classes equally, this result indicates improved class-balanced recognition rather than gains concentrated on frequent classes. In addition, the convolution-based variant ROOTS (ConvNeXt-Base) achieves the highest Top-5 accuracy (99.23%), supporting the robustness of the hierarchy-aware design.

To further examine model robustness under long-tailed conditions, we partition the dataset by class frequency into head classes (top 30%), medium-frequency classes (middle 50%), and tail classes (bottom 20%). Table 2 reports subset-wise Top-1 accuracy under this partition protocol. Traditional purely visual baselines, such as the

Table 2 Top-1 accuracy across different frequency ranges on the mixed dataset. The **best** and second-best results in each column are highlighted. Head/Medium/Tail denote the top 30%/middle 50%/bottom 20% classes by frequency.

Method	Head (%)	Medium (%)	Tail (%)
<i>Pure Vision (Scratch)</i>			
ConvNeXt-Base	93.39	91.65	82.86
ResNet-101	93.28	91.71	82.70
ResNet-152	93.19	91.57	82.91
SwinV2-Base	94.34	93.23	85.76
<i>Pure Vision (Pre-trained)</i>			
ConvNeXt-Base	<u>94.63</u>	93.32	85.92
ResNet-101	92.96	91.16	82.21
ResNet-152	93.16	91.16	82.21
SwinV2-Base	94.34	93.23	85.76
ROOTS (ConvNeXt-Base)	93.45	<u>95.17</u>	<u>91.23</u>
<i>VLM-based</i>			
ROOTS (Qwen3-VL-2B)	94.55	99.98	99.83

ResNet family and SwinV2, show substantial degradation on sparse tail classes, often dropping from 93%–94% on head classes to around 82%–85% on tail classes. By contrast, ROOTS (ConvNeXt-Base) improves tail accuracy from the strongest baseline level of 85.92% to 91.23%. The VLM-based variant ROOTS (Qwen3-VL-2B) yields the largest gain on medium- and tail-frequency classes, including 99.98% accuracy on the medium-frequency subset under the current evaluation protocol and 99.83% on the tail subset.

Figure 9 shows detailed results on individual subsets. By comparing Fig. 9 (a) and Fig. 9 (b), the performance of different models on five datasets (HUST-OBC, OBC306, OBI125, Oracle-50K-125, and PictOBI-20k) exhibits a clear ranking pattern. The red star-marked curve corresponding to ROOTS (Qwen3-VL-2B) consistently leads other methods in both Top-1 and Top-5 accuracy. The yellow solid curve corresponding to ROOTS (ConvNeXt-Base) ranks second and consistently outperforms purely visual baselines, including the ResNet family, SwinV2, and the original ConvNeXt. Overall, ROOTS substantially reduces the performance gap across subsets from different sources.

Figure 10 provides case-level evidence for long-tailed oracle bone character recognition and is consistent with the quantitative results above. The figure presents representative examples from high-frequency, medium-frequency, and tail categories. Regardless of category frequency, the model ranks the ground-truth class at Top-1, while Top-2 and Top-3 are typically visually similar distractors. This behavior suggests that the model is not driven only by class-frequency bias and instead learns fine-grained structural discrimination.

This effect is particularly important for long-tailed classes, where only a few samples are available and glyphs are easily confused with visually similar characters. ROOTS can still identify the target class by leveraging the structural, detailed, and

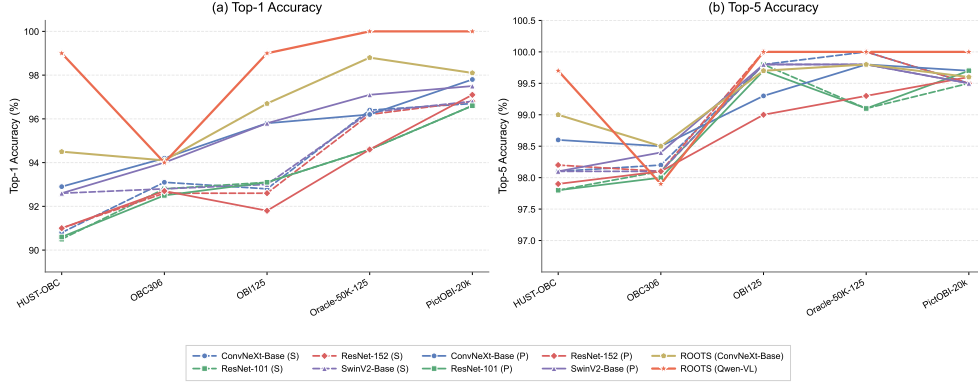


Fig. 9 Per-subset accuracy comparison across five oracle bone benchmarks. (a) Top-1 accuracy and (b) Top-5 accuracy are plotted for each test subset (HUST-OBC, OBC306, OBI125, Oracle-50K-125, PictOBI-20k). Methods are trained either from scratch (S) or with ImageNet pretraining (P). ROOTS (Qwen-VL), marked by red stars, consistently achieves the highest Top-1 accuracy across all subsets, reaching 100% on OBI125, Oracle-50K-125, and PictOBI-20k. ROOTS (ConvNeXt-Base), shown in yellow, ranks second overall and attains the highest Top-5 accuracy among purely visual methods. Both ROOTS variants substantially reduce performance variance across heterogeneous data sources compared to baseline models.

characteristic information encoded in candidate descriptions. This suggests that the hierarchical recognition mechanism, which combines superclass priors with fine-grained discrimination, effectively alleviates the tendency of conventional models to misclassify tail classes as head classes.

In addition, the incorrect candidates shown in Fig. 10 usually share genuine similarities with the ground-truth label in overall contour or local stroke composition. This implies that the text prompts constructed in this work are not simple label annotations; instead, they provide an interpretable basis for decision making, allowing the model to calibrate visual evidence against semantic priors in a consistent manner. Therefore, this visualization not only provides intuitive support for the substantial gains of ROOTS on Macro Top-1 and tail accuracy, but also shows that the proposed method improves both final recognition performance and the interpretable modeling of oracle bone structural knowledge.

Ablation Study

To assess the contribution of each component in the proposed ROOTS framework, we conduct a stepwise ablation study using Qwen3-VL-2B as the base model [32]. The analysis focuses on four factors: mixed-source data construction, domain corpus injection, ROOTS-based knowledge organization, and scaling to a larger vision-language model.

As reported in Table 3, the performance improves consistently as these components are introduced, suggesting that the effectiveness of ROOTS arises from the joint contribution of data enrichment, domain knowledge, structured semantics, and model capacity.

Freq.	Input	GT	Top-1	Top-2	Top-3	Candidate Descriptions (Text Prompt)
High-freq $n = 617$		丁	Correct 丁	Wrong 辛	Wrong 戊	【丁】结构: 四边全封闭, 近似矩形 细节: 转角圆润, 无内部笔画 特征: 空心闭合四边形轮廓
Mid-freq $n = 25$		丙	Correct 丙	Wrong 彭	Wrong 戊	【丙】结构: 拱形对称, 下端开放, 重心偏上 细节: 顶角内侧有两道相向短斜线 特征: 顶角下有内向斜撑结构 【戊】结构: 左右组合, 右部半包围 细节: 左侧人形, 右顶交叉符 特征: 人形附于带叉的门形 【册】结构: 多竖笔被横笔贯穿 细节: 右侧横笔合并呈钩状 特征: 右侧收束的交叉网格
Long-tail $n = 1$		豕	Correct 豕	Wrong 彘	Wrong 彘	【豕】结构: 中轴贯通, 多层横向交叉 细节: 笔画末端尖锐, 具双横杠 特征: 顶部八字形, 下接双横杆 【豕】结构: 竖轴为核, 左侧分叉, 开放结构 细节: 顶部交叉, 左侧双横, 下端尖锐 特征: 竖轴左侧具双层横向笔画
Long-tail $n = 1$		姆	Correct 姆	Wrong 查	Wrong 姥	【姆】结构: 上下组合, 上分叉下交叉 细节: 顶部枝干分叉, 下部线条交织 特征: 下方为多重交叉网状 【姥】结构: 上下分体, 下部对称, 重心偏下 细节: 上部为星芒状, 中部线条交叉 特征: 对称基座上置星芒符号 【姆】结构: 上下组合, 上部三竖分离, 下部人形 细节: 头部呈圈状, 单腿下跪姿态 特征: 人形顶负三道竖笔
Long-tail $n = 1$		帛	Correct 帛	Wrong 彤	Wrong 彊	【帛】结构: 中轴贯穿, 顶部交叉, 下部分叉 细节: 左侧附短斜笔, 下呈人字形 特征: 木形结体, 左侧多一短斜笔 【彤】结构: 右侧主干, 左侧分叉, 开放结构 细节: 左侧含两道平行短斜画 特征: 单侧羽毛形态, 笔画同向 【彊】结构: 左右合体, 右侧分上下 细节: 左为火焰形, 右上为目形 特征: 火旁跪人, 以目为首

Fig. 10 Case-level visualization of ROOTS predictions across frequency categories. Each row represents a test sample from high-frequency ($n = 617$), mid-frequency ($n = 25$), or long-tail ($n = 1$) classes, where n denotes the number of training samples. Columns show: input image, ground-truth label (GT), Top-1/2/3 predictions, and candidate semantic descriptions. Across all cases, ROOTS correctly ranks the ground-truth class at Top-1, while Top-2 and Top-3 candidates are visually similar distractors. The rightmost column displays structured text prompts for each candidate, organized into three levels: Structure (结构), Detail (细节), and Cue (特征), demonstrating that the model leverages interpretable glyph-level semantics rather than relying solely on class-frequency bias.

Table 3 Stepwise ablation study of ROOTS components. Tail denotes Top-1 accuracy on tail classes. M-Top-1 denotes Macro Top-1 accuracy.

Methods	Tail (%)	Top-1 (%)	M-Top-1 (%)
Qwen3-VL-2B	2.53	7.56	5.62
+MixData	37.34	42.24	39.63
+corpus (train)	50.24	65.73	61.24
+corpus (ROOTS)	99.83	95.28	99.38
+Qwen3-VL-8B	99.98	97.73	99.85

When Qwen3-VL-2B is directly applied to oracle bone character recognition, it achieves 2.53% Tail accuracy, 7.56% Top-1 accuracy, and 5.62% Macro Top-1 accuracy. These results indicate that a general-purpose VLM, without task-specific adaptation, has limited effectiveness on this task.

After introducing MixData, the performance increases to 37.34% Tail accuracy, 42.24% Top-1 accuracy, and 39.63% Macro Top-1 accuracy. This improvement suggests that mixed-source data helps alleviate the distribution mismatch caused by isolated datasets and provides the model with a broader range of glyph forms and contextual variations. The gain on the Tail metric is particularly notable, indicating that expanded data coverage is beneficial for improving recognition of underrepresented categories.

Adding corpus (train) further improves the performance to 50.24% Tail accuracy, 65.73% Top-1 accuracy, and 61.24% Macro Top-1 accuracy. Compared with the MixData setting, this result suggests that visual information alone is still insufficient for robust discrimination of rare classes. Incorporating domain text into training appears to strengthen the alignment between visual representations and oracle-bone-specific semantics, thereby providing additional guidance in low-resource settings and improving both overall and class-balanced performance.

A more pronounced improvement is observed when corpus (train) is replaced with corpus (ROOTS). Under this setting, the model reaches 99.83% Tail accuracy, 95.28% Top-1 accuracy, and 99.38% Macro Top-1 accuracy. This result indicates that the gains cannot be explained solely by adding more textual material. Instead, the organization of oracle bone knowledge in ROOTS appears to play a central role by integrating glyph structure, semantic relations, and category hierarchy into a structured knowledge framework.

The attention comparison for corpus injection is shown in Fig. 11.

Oracle-bone Image	w/o Corpus	w Corpus	Corpus
 豕 To follow/Sacrifice	 豕 Pig	 豕	Structure: a continuous central axis with multiple horizontal crossings. Detail: sharp stroke ends and double horizontal bars. Cue: an eight-shaped top connected to twin horizontal strokes.
 棘 Trough	 棘 brambles	 棘	Structure: a central axis running through the form, with upper-lower symmetry and a three-part layout. Detail: claw-like multi-branch ends on both sides, with a cross in the middle. Cue: a closed pouch-like middle section containing a cross.
 楚 To capture	 至 To arrive	 楚	Structure: stacked upper and lower parts with central symmetry and a centered balance. Detail: forked double horns at the top and a diamond-shaped middle block. Cue: a vertical diamond-shaped closed loop in the center.

Fig. 11 Effect of corpus injection on model attention. Each row shows an oracle bone character with its attention maps visualized via Grad-CAM under two conditions: without corpus guidance (w/o Corpus) and with corpus guidance (w Corpus). Warm colors indicate high attention intensity. Row 1 (the OBC representing “to follow/sacrifice”): Without corpus, attention disperses and the model misclassifies the input as “pig”; with corpus, attention concentrates on the central axis and horizontal crossings, yielding the correct prediction. Row 2 (the OBC representing “trough”): Corpus injection sharpens focus on the symmetric three-part structure and claw-like branch ends. Row 3 (the OBC representing “to capture”): Without guidance, the model predicts “to arrive”; corpus redirects attention to the forked double horns and diamond-shaped center, correcting the prediction. The rightmost column displays the injected semantic descriptions organized into Structure, Detail, and Cue, demonstrating how domain knowledge guides visual grounding toward discriminative glyph features.

Figure 11 presents a qualitative comparison of the influence of corpus prior on the attention distribution of Qwen-VL for oracle bone character recognition. Without prior guidance (w/o Corpus), the model exhibits diffuse and incoherent attention patterns and is easily distracted by locally similar features, which may lead to misclassification, for example, predicting the OBC representing “to capture” as “to arrive”. After introducing corpus descriptions containing structural, detailed, and clue-level information (with Corpus), the attention maps become more concentrated on discriminative glyph regions, such as the forked double-horn structure or the central vertical axis, leading to substantially better alignment between visual feature extraction and domain-specific knowledge. This transition from random scattering to semantically focused attention provides strong evidence that external prior knowledge can effectively guide visual grounding in vision-language models under low-resource and highly ambiguous conditions, thereby correcting prediction bias while improving both the robustness and interpretability of model decisions.

Finally, replacing Qwen3-VL-2B with the larger Qwen3-VL-8B further improves the performance to 99.98% Tail accuracy, 97.73% Top-1 accuracy, and 99.85% Macro Top-1 accuracy. Although this improvement is smaller than that obtained from introducing corpus (ROOTS), the result still suggests that increased model capacity provides additional benefits once domain alignment and structured knowledge support are already in place.

Overall, the ablation study supports four observations. First, mixed-source data improves domain generalization by broadening glyph coverage and reducing distribution bias. Second, domain corpus injection enhances the alignment between visual features and oracle-bone-specific semantics. Third, the ROOTS knowledge framework contributes substantially to the recognition of tail classes and to balanced performance across categories. Fourth, a larger VLM further improves the final performance, although its contribution is more limited than that of structured domain knowledge. Taken together, these results suggest that for oracle bone character recognition under long-tailed distributions, improvements cannot be attributed to model scale alone; rather, they rely on the combination of data diversity, semantic supervision, and structured knowledge organization.

Discussion

The experimental results indicate that the proposed design is effective. The purely visual ROOTS variant improves class-balanced recognition over strong pre-trained baselines, especially on sparse tail categories. The multimodal ROOTS variant further shows that integrating structured semantic descriptions with visual evidence improves robustness under long-tailed conditions. In particular, the gains in Macro Top-1 accuracy and tail-category accuracy indicate that ROOTS strengthens discrimination for rare and easily confused characters, which is the key bottleneck in practical oracle bone character recognition.

Case-level visualization and ablation results suggest that the gains are not explained by model scale alone, but by the interaction among mixed-source data construction, domain-specific semantic supervision, hierarchical knowledge organization,

and multimodal inference. More broadly, these findings indicate that oracle bone character recognition benefits from tighter integration of computer vision, multimodal learning, and paleographic structural knowledge.

Despite these gains, three limitations remain. First, the quality of superclass construction still depends on prototype-feature quality and clustering hyperparameters. Second, the semantic knowledge base is partially generated by a general-purpose model and may contain noisy descriptions. Third, evaluation in this study is conducted on mixed public datasets; further validation on newly excavated materials and heavily damaged rubbings is still needed.

Conclusion

This study presents ROOTS, a hierarchical and knowledge-enhanced framework for long-tailed oracle bone character recognition. Unlike prior approaches that mainly mitigate data scarcity through augmentation or optimize recognition directly within the original imbalanced label space, ROOTS addresses the problem by reorganizing the class system itself. Through deep-feature-based prototype extraction and size-aware hierarchical clustering, ROOTS constructs a balanced superclass structure that reduces the difficulty of fine-grained recognition while preserving meaningful glyph relationships. The resulting coarse-to-fine pipeline is aligned with expert analysis workflows and improves class-balanced recognition in both visual-only and VLM-based settings. Overall, ROOTS provides an effective, interpretable, and extensible solution for long-tailed oracle bone character recognition and offers a practical foundation for AI-assisted research on ancient Chinese writing systems.

Future work will focus on three directions: incorporating stronger paleographic supervision for superclass construction, enriching the semantic knowledge base with contextual and philological information, and validating the framework on more diverse excavation materials and severely damaged rubbings.

Acknowledgements. This work was supported in part by the National Social Science Fund of China under Grant 24BYY005, in part by Shanghai Jiao Tong University’s Key Project on Intelligent Humanities under Grant ZHWK2506, and in part by the Ministry of Education’s “Chunhui Plan” Cooperative Scientific Research Project, “Research on Intelligent Detection and Recognition Models and Applications for Oracle Bone Inscriptions,” under Grant HZKY20220067.

Author contributions. Ziqi Li and Zijian Chen initiated this research and gradually developed the overall framework. Tingzhu Chen and Ziyi Yang, as oracle bone inscription experts, contributed to the construction of standardized data and corpora, and helped refine the model-training procedure. Ziqi Li completed the code implementation and the initial draft of the manuscript, which was thoroughly proofread and reviewed by Zijian Chen, Tingzhu Chen, and Ziyi Yang. Tingzhu Chen and Guangtao Zhai conceived and supervised the project. All authors discussed the final manuscript and agreed to submit it for publication.

Competing interests. The authors declare no competing interests.

Additional information. Correspondence and requests for materials should be addressed to Tingzhu Chen.

References

- [1] Chongsheng Zhang, Ruixing Zong, Shuang Cao, Yi Men, and Bofeng Mo. AI-Powered Oracle Bone Inscriptions Recognition and Fragments Rejoining. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5309–5311, 2020. doi: 10.24963/ijcai.2020/779.
- [2] Haisu Guan, Huanxin Yang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. Deciphering Oracle Bone Language with Diffusion Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15554–15567, 2024. doi: 10.18653/v1/2024.acl-long.831.
- [3] Zijian Chen, Tingzhu Chen, Wenjun Zhang, and Guangtao Zhai. OBI-Bench: Can LMMs Aid in Study of Ancient Script on Oracle Bones? In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://openreview.net/forum?id=hL5jone2Oh>.
- [4] Zijian Chen, Yuan Tian, Yuze Sun, Wei Sun, Zicheng Zhang, Weisi Lin, Guangtao Zhai, and Wenjun Zhang. Just noticeable difference for large multimodal models. *arXiv preprint arXiv:2507.00490*, 2025.
- [5] Zijian Chen, Wenjie Hua, Jinhao Li, Yucheng Zhu, Xiaona Zhi, Zhiji Liu, Tingzhu Chen, Wenjun Zhang, and Guangtao Zhai. Oracle bone inscriptions information processing: a comprehensive survey. *npj Heritage Science*, 14:220, 2026. doi: 10.1038/s40494-026-02511-w.
- [6] Lu Yang, He Jiang, Qing Song, and Jun Guo. A Survey on Long-Tailed Visual Recognition. *International Journal of Computer Vision*, 130:1837–1872, 2022. doi: 10.1007/s11263-022-01622-8.
- [7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-Balanced Loss Based on Effective Number of Samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269, 2019. doi: 10.1109/CVPR.2019.00949.
- [8] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling Representation and Classifier for Long-Tailed Recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. URL <https://openreview.net/forum?id=r1gRTCvFvB>.
- [9] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In

- [10] Wang W., Zhang T., Zhao Y., Jin X., Mouchère H., and Yu X. Improving oracle bone characters recognition via a cyclegan-based data augmentation method. In *International Conference on Neural Information Processing*, pages 88–100, Singapore, 2022. Springer Nature Singapore. doi: 10.1007/978-981-99-1645-0_8.
- [11] Jiaying Gao, Fausto Giunchiglia, Tongyu Zhao, Xin Zhao, Chuntao Li, and Hao Xu. Oracle bone heritage data augmentation based on two-stage decomposition GANs. *npj Heritage Science*, 13:230, 2025. doi: 10.1038/s40494-025-01774-z.
- [12] Xuebin Yue, Hengyi Li, Yoshiyuki Fujikawa, and Lin Meng. Dynamic Dataset Augmentation for Deep Learning-based Oracle Bone Inscriptions Recognition. *Journal on Computing and Cultural Heritage*, 15(4):1–20, 2022. doi: 10.1145/3532868.
- [13] Jing Li, Qiu-Feng Wang, Kaizhu Huang, Xi Yang, Rui Zhang, and John Y. Goulermas. Towards better long-tailed oracle character recognition with adversarial data augmentation. *Pattern Recognition*, 140:109534, 2023. doi: 10.1016/j.patcog.2023.109534.
- [14] Yue Zhang, Aoli Han, Shibin Wang, Jie Liu, Xueshan Li, Dong Liu, and Yigang Cen. Generating oracle bone inscriptions based on the structure aware diffusion model. *npj Heritage Science*, 13:461, 2025. doi: 10.1038/s40494-025-02000-6.
- [15] Jinhao Li, Zijian Chen, Runze Jiang, Tingzhu Chen, Changbo Wang, and Guangtao Zhai. Mitigating Long-tail Distribution in Oracle Bone Inscriptions: Dataset, Model, and Benchmark. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7729–7738, 2025. doi: 10.1145/3746027.3755067.
- [16] Junwei Zhou, Qingpeng Tu, and Guojun Xu. Oracle character recognition using universal inverted bottleneck and inverse image frequency. *International Journal on Document Analysis and Recognition*, 28(4):609–622, 2025. doi: 10.1007/s10032-025-00512-2.
- [17] Zhengchen Li, Xiuan Wan, Jiahua Wu, Shouyong Pan, and Yuchun Fang. Historical oracle bone character recognition through domain transfer and mutual learning. *Applied Soft Computing*, 175:113031, 2025. doi: 10.1016/j.asoc.2025.113031.
- [18] Carlos N. Silla Jr and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22: 31–72, 2011. doi: 10.1007/s10618-010-0175-9.

- [19] Yanyun Qu, Li Lin, Fumin Shen, Chang Lu, Yang Wu, Yuan Xie, and Dacheng Tao. Joint Hierarchical Category Structure Learning and Large-Scale Image Classification. *IEEE Transactions on Image Processing*, 26(9):4331–4346, 2017. doi: 10.1109/TIP.2016.2615423.
- [20] V. M. Araújo, A. S. Britto Jr, L. S. Oliveira, and A. L. Koerich. Two-view fine-grained classification of plant species. *Neurocomputing*, 467:427–441, 2022. doi: 10.1016/j.neucom.2021.10.015.
- [21] Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. HDLTex: Hierarchical Deep Learning for Text Classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371, 2017. doi: 10.1109/ICMLA.2017.0-134.
- [22] Yaowei Yang, Zhonghua Sun, Kaisaier Tuerxun, and Kurban Ubul. Scene Script Identification Using Dense Hierarchical Semantic Fusion. In *Document Analysis and Recognition – ICDAR 2025: 19th International Conference, Wuhan, China, September 16–21, 2025, Proceedings, Part III*, pages 358–374, 2025. doi: 10.1007/978-3-032-04624-6_21.
- [23] Rohit Babbar, Ioannis Partalas, Eric Gaussier, and Massih R. Amini. On Flat versus Hierarchical Classification in Large-Scale Taxonomies. In *Advances in Neural Information Processing Systems 26*, pages 1824–1832, 2013.
- [24] M. Oquab, T. Darcet, T. Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [25] Joe H. Ward Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845.
- [26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022. doi: 10.1109/CVPR52688.2022.01167.
- [27] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142, 2023. doi: 10.1109/CVPR52729.2023.01548.
- [28] Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Jinpeng Wan, Haisu Guan, Zhebin Kuang, Lianwen Jin, Xiang Bai, and Yuliang Liu. An Open Dataset for Oracle Bone Character Recognition and Decipherment. *Scientific Data*, 11:976, 2024. doi: 10.1038/s41597-024-03807-x.

- [29] S. Huang, H. Wang, Y. Liu, X. Shi, and L. Jin. Obc306: A large-scale oracle bone character recognition dataset. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 681–688, 2019. doi: 10.1109/ICDAR.2019.00114.
- [30] Zijian Chen, Wenjie Hua, Jinhao Li, Lirong Deng, Fan Du, Tingzhu Chen, and Guangtao Zhai. PictOBI-20k: Unveiling Large Multimodal Models in Visual Decipherment for Pictographic Oracle Bone Characters, 2025. arXiv preprint arXiv:2509.05773.
- [31] Bang Li, Jing Yang, Yujie Liang, Xiaobin Hu, Zengmao Ding, Xu Peng, Shengwei Han, Peichao Qin, Donghao Luo, Taisong Jin, Feng Gao, Yongge Liu, and Rongrong Ji. OBIMD: A Multi-modal Dataset for Contextual Interpretation of Oracle Bone Inscriptions. *Scientific Data*, 2026. doi: 10.1038/s41597-026-06967-0. Article in Press.
- [32] S. Bai, Y. Cai, R. Chen, et al. Qwen3-vl technical report, 2025. arXiv preprint arXiv:2511.21631.
- [33] Caoshuo Li, Zengmao Ding, Xiaobin Hu, Bang Li, Donghao Luo, AndyPian Wu, Chaoyang Wang, Chengjie Wang, Taisong Jin, Seven Shu, Yunsheng Wu, Yongge Liu, and Rongrong Ji. OracleFusion: Assisting the Decipherment of Oracle Bone Script with Structurally Constrained Semantic Typography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19893–19902, 2025. URL <https://openaccess.thecvf.com/content/ICCV2025/html/Li-OracleFusion-Assisting-the-Decipherment-of-Oracle-Bone-Script-with-Structurally-ICCV-2025-paper.html>.
- [34] Jing Li, Bin Dong, Qiu-Feng Wang, Lei Ding, Rui Zhang, and Kaizhu Huang. Decoupled Learning for Long-Tailed Oracle Character Recognition. In *Document Analysis and Recognition – ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part IV*, pages 165–181. Springer, 2023. doi: 10.1007/978-3-031-41685-9_11.
- [35] ZhongYuan Yang, Ghalip Ibrahim, Xuebin Xu, Nurbiya Yadikar, and Kurban Ubul. Long-tailed Oracle Recognition Based on Attention and Multi-Expert Learning. In *2024 IEEE 5th International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 408–413, 2024. doi: 10.1109/PRML62565.2024.10779624.
- [36] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing Network Design Spaces. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10425–10433, 2020. doi: 10.1109/CVPR42600.2020.01044.
- [37] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective Kernel Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

- (*CVPR*), pages 510–519, 2019. doi: 10.1109/CVPR.2019.00060.
- [38] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip H. S. Torr. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2021. doi: 10.1109/TPAMI.2019.2938758.
- [39] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009, 2022. doi: 10.1109/CVPR52688.2022.01170.