Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond Aggregate Metrics Diagnosing the Long-Tailed Failure of Oracle Bone Character Detection and Recognition

Dan Sui
Anyang Normal University
Yuhang Xing
Anyang Normal University
Penglan Liu
University of Jyväskylä
Bang Li
libang@aynu.edu.cn

Anyang Normal University

Research Article

Keywords: oracle bone inscriptions, character detection, character recognition, long-tailed recognition, few-shot recognition, YOLO, OBIMD, digital humanities

Posted Date: July 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10039997/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 Discover Artificial Intelligence on October 12th, 2026. See the published version at https://doi.org/10.1007/s44163-026-02210-2.

Beyond Aggregate Metrics: Diagnosing the Long-Tailed Failure of Oracle Bone Character Detection and Recognition

Dan Sui[1,2†], Yuhang Xing[1†], Penglan Liu[3], Bang Li[2*]

[1]School of Computer and Information Engineering, Anyang Normal University, Anyang, 455000, China.
[2]Key Laboratory of Oracle Bone Inscriptions Information Processing, Ministry of Education, Anyang, 455000, China.
[3]University of Jyväskylä, Jyväskylä, 40700, Finland.

*Corresponding author(s). E-mail(s): libang@aynu.edu.cn;
†These authors contributed equally to this work.

Abstract

Oracle bone inscriptions, the earliest mature Chinese writing system, are a primary source for studying ancient Chinese civilization, yet their interpretation remains labor-intensive. Automatically detecting and recognizing visible characters on rubbing images is therefore an entrance task for AI-assisted interpretation. The Oracle Bone Inscription Multi-modal Dataset exposes the problem of extreme long-tailed class distribution: aggregate evaluation conceals severe performance defects on rare classes — the detector can reach an overall AP_{50} of 0.63, while the few-shot class AP_{50} is only 0.06. we conduct a systematic diagnosis with a frequency-aware stratified evaluation protocol, confirming that the task bottleneck lies in long-tail recognition rather than class-agnostic localization, and naive reweighting within the detection head is counterproductive. We thus propose FAR-YOLO, a minimal decoupled baseline that separates localization and recognition training, optimizing the recognition branch only with frequency-aware sampling and morphology-preserving augmentation. While maintaining nearly unchanged overall AP_{50}, FAR-YOLO improves the few-shot AP_{50} by 10.5 percentage points.

Keywords: oracle bone inscriptions, character detection, character recognition, long-tailed recognition, few-shot recognition, YOLO, OBIMD, digital humanities

1 Introduction

Oracle bone inscriptions are the earliest known mature Chinese writing system, inscribed on turtle shells and animal bones more than three thousand years ago, and they constitute primary evidence for the study of early Chinese language, divination, ritual, and social organization. Their interpretation, however, remains slow and dependent on a small community of specialists, because the excavated materials are fragmentary, visually degraded, and written in highly variable ancient glyph forms. Computational methods can support paleographic research by enabling large-scale indexing, variant comparison, and contextual retrieval, but these higher-level functions all rely on a more basic visual capability: a system must first read the visible characters in an oracle bone image. For rubbing images, one of the most widely used visual records of oracle bone materials, this requires both locating each visible character and assigning it a glyph identity. Oracle bone character detection and recognition therefore form an entrance task for AI-assisted interpretation of oracle bone inscriptions.

Research on this task has been strongly shaped by the available annotations. Early detectors based on SSD, YOLO, and Faster R-CNN, together with the OBI detection dataset (later referred to as ODD), established character localization from rubbing images as a measurable benchmark [1, 2]. Because these resources provide character regions but not large-scale glyph identities, research could reliably evaluate whether a character was found, but not which glyph it was, and the field therefore concentrated on localization rather than complete detection and recognition. A few studies moved beyond pure localization, but their recognition was restricted to small category sets or to recognizers trained on separate datasets, such as the two-stage Prism-OBI with its 306-class cropped recognizer [8]; none defined a large-scale full-image setting in which localization and glyph recognition are supervised and evaluated on the same dense annotations.

The release of the Oracle Bone Inscription Multi-modal Dataset (later referred to as OBIMD) [9], which provides both bounding boxes and glyph categories on the same rubbing images, makes large-scale full-image oracle bone character detection and recognition possible for the first time. Among public oracle resources, OBIMD is uniquely suited to this study: earlier detection-recognition datasets cover only tens to a few hundred glyph classes (for example, 40 classes [1] or 306 [8]), too few to exhibit a realistic large-vocabulary long tail, whereas OBIMD spans more than 1,700 categories on full rubbing images. It is therefore the only existing testbed for this problem, and our quantitative conclusions are scoped accordingly. Extending the task to more than one thousand glyph categories, however, reveals an extreme long-tailed distribution: a few frequent categories dominate the training signal, while most categories contain only a handful of samples (Figure 1). This inherits the long-tail and few-shot difficulties already studied in cropped oracle character recognition, where data augmentation, feature decoupling, and frequency-aware learning have been used to improve low-frequency categories [10–13], and it indicates that oracle character recognition is far from a balanced closed-set classification problem.

These recognition methods, however, assume that character crops are already available, and do not consider how long-tailed classification interacts with localization

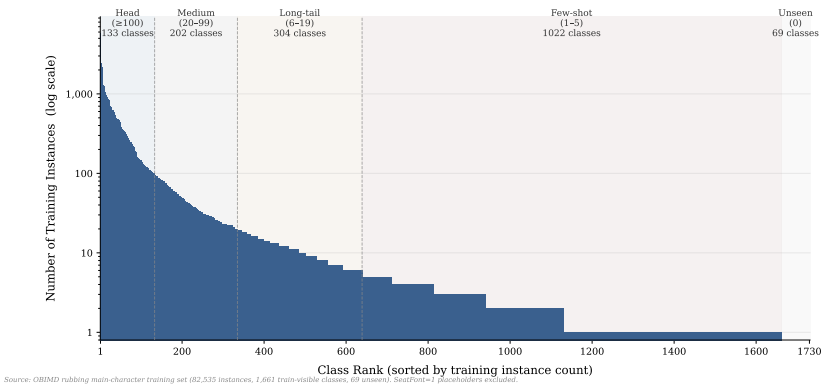


Fig. 1 Frequency distribution of OBIMD rubbing main-character categories. The dataset exhibits an extreme long-tailed distribution, with over 60% of categories belonging to the few-shot regime.

and confidence estimation inside a detector. Under the full-image setting, this interaction produces a failure that conventional evaluation conceals: a competitive direct multi-class YOLO baseline reaches 0.63 overall AP_{50} across three seeds, yet only 0.06 few-shot AP_{50}. Aggregate AP cannot reflect such disparities across frequency groups, which exposes a limitation of the evaluation protocols currently used for this task.

These observations point to a structural hypothesis: under extreme long-tail distributions, the closed-set detector formulation introduces a harmful coupling between localization—which is robust and category-independent—and recognition—which suffers from severe class imbalance. To test this hypothesis, we construct a simple decoupled pipeline, called FAR-YOLO (Frequency-Aware Recognition YOLO), in which a class-agnostic YOLO localizer handles character presence and a separate crop-level recognizer handles glyph identity. The recognizer uses only straightforward few-shot-aware exposure rebalancing and morphology-preserving data augmentation. This design is intentionally minimal: its role is to probe the task structure and set a strong baseline, not to introduce a novel model class. The frequency-aware evaluation protocol makes the resulting rare-category behavior measurable and interpretable.

The contributions of this paper are as follows:

1. We diagnose the task structure of large-scale oracle bone character detection and recognition. Long-tail learning is well studied for pre-cropped oracle characters, but its interaction with localization inside a full-image detector has not been examined. Evaluating on the OBIMD rubbing main-character setting under a frequency-aware protocol, we show that (i) contrary to the field's emphasis on detector engineering, class-agnostic localization is already strong ($AP_{50} = 0.85$, recall $= 0.80$) and far less limiting than recognition; (ii) the obvious in-head remedy backfires, as naive inverse-frequency reweighting of the detector head lowers overall AP_{50} rather than helping the tail; and (iii) a minimal decoupled design substantially recovers few-shot performance yet leaves the tail far from solved, while ranked candidates frequently contain the correct rare glyph.

2. Adapting the stratified evaluation practice common in long-tailed detection, we report a frequency-aware breakdown that partitions categories into head, medium, long-tail, few-shot, and unseen groups. Per-frequency AP is itself standard; applying it to this task is what exposes the rare-class collapse that aggregate AP conceals, and we release the evaluation suite so that such diagnosis is reproducible.
3. We provide FAR-YOLO, a decoupled baseline that pairs a class-agnostic YOLO localizer with a frequency-aware crop-level recognizer. FAR-YOLO improves few-shot AP_{50} by 10.5 points and class-balanced AP_{50} by 3.7 points over a direct multi-class detector while maintaining overall AP_{50}, without aggressive classifier replacement, complex reweighting, or extra data.
4. practical oracle bone character reading should move beyond closed-set forced classification toward candidate ranking, retrieval-assisted recognition, and paleographer-in-the-loop verification.

2 Related Work

2.1 Oracle Character Detection

Early studies have validated the effectiveness of detectors such as SSD, YOLO, and Faster R-CNN for oracle character localization from rubbing images [1, 14, 15], ODD established a shared localization benchmark [2], enabling systematic comparison of detectors under complex backgrounds, dense layouts, and broken strokes. Subsequent work further improved localization performance through multi-scale feature enhancement [16], small-object adaptation [17–21], and cross-domain robustness [22, 23]. Recent YOLO-based variants have also introduced multi-branch feature fusion and efficient attention mechanisms to improve accuracy [24]. However, existing detection datasets and evaluation protocols only focus on the accuracy of character region localization, without providing dense glyph category annotations.

2.2 Oracle Character Detection and Recognition

Several studies have introduced recognition information beyond pure localization. Early SSD experiments covered only tens of character classes on a small number of rubbing images [1], while engineering systems such as YOLO combined with MobileNet relied on self-built small datasets [3]. Other work adopted a two-step approach of "single-class detection followed by cropped character classification," but the detection annotations contained only position information without glyph category labels [4]. Furthermore, prototype-guided cyclic disentanglement, shape prior fusion, and pseudo-category labeling, although incorporating glyph information into the detection pipeline [5–7], did not perform end-to-end detection-recognition evaluation based on a unified dense annotation system. Prism-OBI decoupled localization from 306-class cropped character recognition [8], but its recognition vocabulary is far from sufficient to cover the thousand-class glyph space of OBIMD.

Existing detection-recognition studies are either constrained by small category sets or small data scales, or have not been validated under the "full-image dense annotation

with thousands of glyph classes" setting, and therefore cannot address the coupling problem between localization and long-tailed glyph recognition in large-scale scenarios.

2.3 Long-Tailed Oracle Character Recognition

Oracle character recognition inherently exhibits long-tailed distribution characteristics. Existing studies have proposed various solutions for the long-tail and few-shot problems in cropped character recognition: Mix-Up augmentation and TailMix improve minority-class performance through data augmentation [10, 11], decoupled learning separates representation learning from classifier optimization [12], and the FFD augmentor generates few-shot character variants through stroke-level deformation [13]. These methods also draw on core ideas from general long-tailed recognition, such as Focal Loss and Balanced Softmax [25–28]. Broader surveys of oracle character recognition identify data scarcity and writing variability as persistent obstacles [29], while few-shot and prototype-based learning provide a foundation for extremely rare categories [30, 31].

The above studies all assume that "character crop regions have already been accurately obtained," and do not consider the coupled challenges of full-image detection-recognition. A direct YOLO-based multi-class detector must optimize small-object localization, object confidence, and thousand-class long-tailed category boundaries within the same detection head. Frequent categories tend to dominate gradient updates, while the performance of low-frequency and few-shot categories is masked. To date, no research has systematically analyzed the suitability of direct YOLO-style detection-recognition under extreme long-tail conditions, nor has any evaluation protocol been proposed to reveal failures that may be hidden by aggregate metrics.

3 Diagnostic Framework and Decoupled Baseline

3.1 Problem Formulation and Design Rationale

Given a rubbing image x, the full-image detection-recognition task contains a set of visible character annotations $\mathcal{A} = \{(b_i, y_i)\}_{i=1}^{N}$, where b_i is the bounding box of the i-th visible character and y_i is its glyph category. The closed-set recognition space contains the seen categories $\mathcal{C}_s$, while categories without training samples are retained only for unseen/open-set diagnosis. A prediction is correct only when both localization and glyph identity are correct, so the task is stricter than single-class character localization.

A direct multi-class YOLO detector represents each candidate with a box, an objectness score, and a $|\mathcal{C}_s|$-way glyph posterior. Its training objective can be written abstractly as

$$\mathcal{L}_{\text{direct}} = \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{obj}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^{|\mathcal{C}_s|}. \tag{1}$$

This formulation is compact, but it forces box regression, foreground confidence, and more than one thousand imbalanced glyph boundaries to be optimized inside the same dense prediction head. If n_c denotes the number of training samples in class c, the classification contribution of frequent classes is naturally much larger than that of rare classes. Few-shot glyphs therefore provide limited evidence for stable

decision boundaries, while simple rare-class reweighting may perturb the objectness and ranking behavior of the detector head. The decoupled baseline is designed to retain end-to-end prediction at inference time while separating the learning problems of visible-character localization and long-tailed glyph recognition during training. The goal is not to claim architectural novelty, but to instrument a minimal decomposition whose diagnostic value is determined by the experiments.

3.2 Decoupled Localization and Recognition

This diagnostic formulation full-image oracle character detection and recognition into class-agnostic localization followed by crop-level glyph recognition:

$$p(b, y \mid x) \propto p_\theta(b, \text{object} \mid x)\; p_\phi(y \mid \mathbf{z}(b, x)), \tag{2}$$

where p_θ is the localization branch, $\mathbf{z}(b, x)$ is the normalized crop extracted from box b, and p_ϕ is the glyph recognizer. The localization branch is a YOLO11s detector trained with a single foreground class, "oracle character." Its objective removes the dense glyph classification term:

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{obj}} + \lambda_{\text{fg}}\mathcal{L}_{\text{fg}}. \tag{3}$$

All visible characters contribute to foreground localization regardless of glyph identity. This design uses the statistical commonality of the task: although oracle glyph identities are extremely imbalanced, the visual concept of a visible character region is shared across all categories. The detector is therefore optimized for candidate coverage and foreground-background separation rather than for sparse fine-grained class boundaries.

For each predicted box $\hat{b}_j$ with localization confidence q_j, the pipeline crops the corresponding image region, pads it to a square white canvas, resizes it to the recognizer input size, and obtains a glyph posterior $p_\phi(c \mid \mathbf{z}_j)$ over the seen categories. The final class-specific score is

$$s_{j,c} = q_j\; p_\phi(c \mid \mathbf{z}_j), \quad c \in \mathcal{C}_s. \tag{4}$$

The predicted glyph is $\arg\max_c\, p_\phi(c \mid \mathbf{z}_j)$, and the ranked detection-recognition confidence is $\max_c\, s_{j,c}$. This multiplicative fusion keeps localization and recognition uncertainty in the same candidate ranking: a poorly localized crop cannot be rescued solely by a confident recognizer, and a visually plausible box is not ranked highly unless its glyph identity is also confident. From the perspective of direct YOLO, The decoupled baseline replaces the long-tailed dense classification head with a crop-level recognition head whose training distribution can be controlled independently.

Figure 1 illustrates the complete inference and training logic of FAR-YOLO. The upper stream starts from a complete rubbing image and uses class-agnostic YOLO to produce visible-character proposals. Each proposal is converted into a normalized crop and passed to the recognizer, which outputs glyph probabilities over the seen categories. The detector confidence and recognizer confidence are then fused to generate ranked character candidates. The lower stream describes the recognizer training process: crop-level samples are grouped by training frequency, rare classes receive

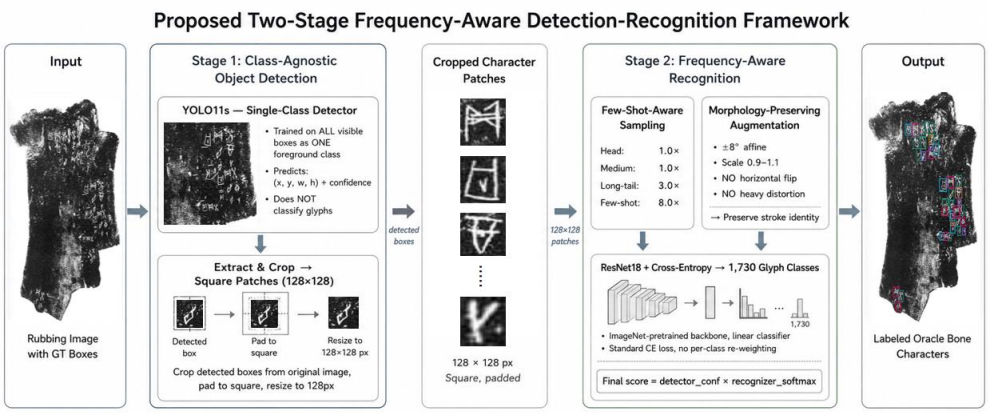


Fig. 2 Overview of the decoupled baseline (FAR-YOLO). The figure shows the full rubbing image, class-agnostic YOLO localization, crop construction, frequency-aware ResNet18 glyph recognition, exposure rebalancing, morphology-preserving augmentation, confidence fusion, confidence-gated acceptance/rejection, and final textual candidates.

higher expected exposure, and morphology-preserving augmentation is applied before the recognition loss. The framework therefore separates the two sources of difficulty in the task: dense small-character localization is handled by the detector, while long-tailed glyph discrimination and uncertain prediction handling are handled after crop extraction.

3.3 Confidence-Gated Recognition and Rejection

In practical oracle character interpretation, forcing every detected crop to a closed-set glyph label is undesirable. Some crops may be poorly localized, severely degraded, ambiguous between variants, or belong to categories absent from the training set. The baseline therefore uses a confidence-gated output rule in addition to ranked AP evaluation. For a detected crop j, the accepted glyph label and confidence are

$$\hat{y}_j = \arg\max_c s_{j,c}, \quad S_j = \max_c s_{j,c}. \tag{5}$$

Given an operating threshold τ, the final output is

$$\text{output}_j = \begin{cases} \hat{y}_j, & \text{if } S_j \geq \tau; \\ \text{reject}, & \text{otherwise.} \end{cases} \tag{6}$$

The reject decision does not introduce a new learned zero-shot class. It is a conservative abstention mechanism that prevents low-confidence crops from being forced into one of the seen categories. Rejected candidates can be passed to expert verification, retrieval-assisted search, or a future open-set recognition module. This is particularly important for OBIMD-style data because unseen categories are present at evaluation time but have no training examples. The threshold τ can be selected on validation data according to the desired precision-recall trade-off. In the controlled experiments,

AP metrics are computed from ranked candidates, while thresholded precision, recall, and F1 use the same score definition to evaluate the behavior of accepted predictions at an operating point. This abstain-and-defer behavior is the operational counterpart of the candidate-ranking direction that our diagnostics support: rather than forcing a low-confidence crop into a seen label, the system can withhold a decision and route the crop to ranked candidates or expert verification.

3.4 Frequency-Aware Recognition Training

After localization is decoupled from glyph classification, the remaining recognition problem can be trained with a frequency-aware objective. Let $\mathcal{D}_{\text{rec}} = \{(\mathbf{z}_i, y_i)\}$ be the crop-level training set and let n_c denote the number of training crops in class c. Each sample receives a group-dependent exposure weight

$$w_i = \begin{cases} 1, & \text{if } y_i \text{ is head or medium;} \\ \alpha, & \text{if } y_i \text{ is long-tail;} \\ \beta, & \text{if } y_i \text{ is few-shot.} \end{cases} \tag{7}$$

The recognizer is trained by sampling with replacement according to

$$P(i) = w_i \Big/ \sum_j w_j. \tag{8}$$

For an epoch containing M sampled crops, the expected exposure of class c becomes

$$\mathbb{E}_c = M \, \frac{n_c w_c}{\sum_d n_d w_d}. \tag{9}$$

Thus, rare classes are not given new labels or external examples; instead, their expected optimization exposure is increased relative to frequent classes. In the experiments, head and medium classes use weight 1, long-tail classes use $\alpha = 3$, and few-shot classes use $\beta = 8$. This design targets the training imbalance directly while leaving the detector unchanged.

The recognition objective further applies morphology-preserving augmentation. Oracle glyph identity depends on stroke topology and relative component layout, so aggressive transformations such as horizontal flipping, heavy mixup, or strong geometric distortion can alter the semantic identity of a character. the recognizer therefore optimizes

$$\mathcal{L}_{\text{rec}}(\phi) = \mathbb{E}_{i \sim P,\, t \sim \mathcal{T}_{\text{morph}}} \big[-\log p_\phi(y_i \mid t(\mathbf{z}_i)) \big], \tag{10}$$

where $\mathcal{T}_{\text{morph}}$ contains mild affine, perspective, and photometric transformations constrained to preserve the overall glyph structure. The role of this augmentation is different from exposure rebalancing: sampling changes how often rare classes appear during optimization, whereas morphology-preserving augmentation increases crop-level robustness around each observed glyph without creating structurally invalid characters.

The complete training procedure is therefore composed of two coordinated objectives:

$$\min_{\theta} \mathcal{L}_{\mathrm{loc}}(\theta), \quad \min_{\phi} \mathcal{L}_{\mathrm{rec}}(\phi). \tag{11}$$

At inference time, the two branches are composed through the score $s_{j,c}$ defined above. The ablation experiments evaluate whether the observed gains come from decoupling alone, rare-class exposure rebalancing, morphology-preserving augmentation, or their combination.

4 Results and Discussion

4.1 Experimental Details

We follow the same target setting as the OBIMD technical validation [9]. This study focuses on the rubbing modality and the main-character label set. Missing-position placeholder annotations are not treated as visible character detection targets in this setting. Under this setting, the rubbing main-character subset contains 9,124 training images, 953 validation images, and 953 test images. The corresponding annotation counts are 82,535, 8,924, and 8,924. The label space contains 1,730 glyph categories in total. Among them, 1,661 categories have at least one training instance and can be evaluated as closed-set categories. The remaining 69 categories have no training instance but appear in validation and test; they are retained as an unseen/open-set diagnostic group.

Classes are divided according to the number of training samples. Head classes contain at least 100 training instances; medium classes contain 20 to 99 instances; long-tail classes contain 6 to 19 instances; few-shot classes contain 1 to 5 instances; unseen classes contain no training instance. This grouping separates ordinary low-frequency classes from the more severe few-shot and open-set regimes.

Table 1 Frequency distribution of OBIMD rubbing main-character categories.

Group	Classes	Train Inst.	Val Inst.	Test Inst.
Head	133	68,404	7,314	7,314
Medium	202	8,789	953	953
Long-tail	304	3,285	377	377
Few-shot	1,022	2,057	206	206
Unseen	69	0	74	74

The experiments separate the task into four complementary settings. Class-agnostic localization measures whether visible character regions can be found before recognition. Ground-truth crop recognition isolates glyph classification by evaluating recognizers on true boxes. Direct multi-class detection evaluates the compact formulation in which YOLO predicts boxes and glyph categories jointly. Two-stage detection-recognition evaluates whether localization and glyph recognition benefit from separate learning mechanisms. FAR-YOLO belongs to the two-stage setting, while direct multi-class YOLO serves as the compact end-to-end baseline.

The published OBIMD technical validation reports a direct YOLOv11l reference on the same rubbing main-character subtask, with mAP@50 of 0.5141, mAP@50:95 of 0.3045, precision of 0.6386, recall of 0.4271, F1 of 0.5118, and average IoU of 0.7214 [9].

4.2 Compared Methods and Implementation

The main comparison includes both direct and decoupled formulations. The direct multi-class YOLO detector predicts bounding boxes and glyph categories jointly and serves as the compact detection-recognition baseline. To test whether the long-tail problem can be handled within the detector itself, a class-weighted direct YOLO variant is trained with inverse-frequency weights applied to the classification loss. This baseline keeps the same direct architecture and modifies only the detector-head classification objective.

The decoupled baselines use the same class-agnostic YOLO localizer as FAR-YOLO and differ only in the crop-level recognizer. The cross-entropy two-stage cascade uses a standard ResNet18 recognizer trained with cross entropy and provides the reference two-stage baseline. The prototype-balanced cascade replaces the linear classifier with a prototype-style classifier and uses Balanced Softmax, representing a stronger rare-class correction strategy. FAR-YOLO uses the same localizer and ResNet18 backbone, but trains the recognizer with few-shot-aware exposure rebalancing and morphology-preserving augmentation.

Additional ablations isolate the effect of each design choice. The oversampling-only variant applies the group-aware sampler without stronger augmentation, whereas the augmentation-only variant applies morphology-preserving augmentation without frequency-aware sampling. Logit-adjusted cascades calibrate recognition logits by class-frequency priors during inference. Balanced Softmax, prototype classification, and prototype classification with focal loss are used as diagnostic alternatives.

The controlled direct detector, class-weighted direct detector, and class-agnostic detector use YOLO11s pretrained weights [32]. They are trained for 30 epochs with image size 768, batch size 16, SGD optimizer, initial learning rate 0.01, final learning-rate factor 0.01, momentum 0.937, weight decay 0.0005, NMS IoU threshold 0.7, and maximum 300 detections per image. The class-weighted direct detector uses YOLO classification-loss weight power $\mathrm{cls_pw} = 0.5$.

All recognizers use an ImageNet-pretrained ResNet18 backbone, 128×128 input crops, batch size 256, 5 training epochs, AdamW optimizer, learning rate 3×10^{-4}, and weight decay 1×10^{-4}. The prototype recognizer replaces the linear classifier with a cosine-normalized prototype head with scale 30. Balanced Softmax adds log class counts to logits during training. Logit adjustment subtracts α times the log class counts during inference.

End-to-end evaluation uses detector confidence threshold 0.05, detector NMS IoU threshold 0.7, maximum 300 detections per image, and exact label matching at IoU threshold 0.5. For cascades, predictions are ranked by the product of detector confidence and recognizer confidence. AP_{50} is computed from ranked predictions under exact box-and-label matching. $\mathrm{mAP}_{50:95}$ averages AP over IoU thresholds from 0.50 to 0.95 with step size 0.05.

4.3 Evaluation Metrics

We report conventional OBIMD detection metrics, including precision, recall, F1, AP_{50}, AP_{75}, $mAP_{50:95}$, and average IoU where applicable, to maintain comparability with the published technical validation. The frequency-aware end-to-end analysis uses four primary AP_{50} metrics. This frequency-stratified reporting follows established practice in long-tailed detection, where AP is computed separately for rare, common, and frequent classes, as in LVIS [33]; we adopt finer bins (head, medium, long-tail, few-shot) and an explicit unseen group suited to the oracle glyph distribution.

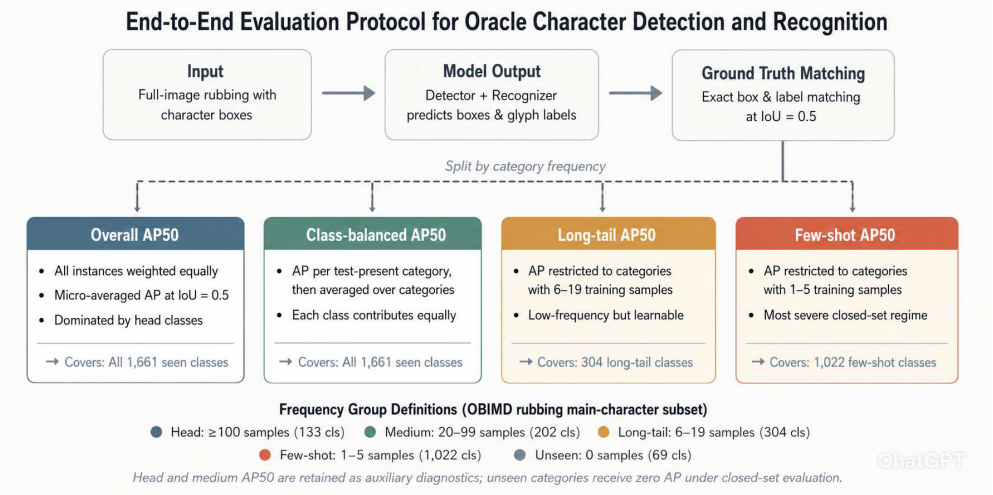


Fig. 3 Evaluation protocol for end-to-end oracle character detection and recognition. The figure distinguishes instance-level overall AP_{50}, class-balanced AP_{50}, long-tail AP_{50}, and few-shot AP_{50}.

Overall AP_{50}, also called micro AP_{50} in the experimental records, evaluates all instances together and reflects instance-dominated detection and recognition performance. Class-balanced AP_{50}, also called macro AP_{50}, first computes AP per test-present category and then averages over categories, reducing the dominance of frequent classes. In the end-to-end closed-set detection protocol, unseen categories are present in the test annotations but cannot be predicted by a recognizer trained only on seen classes; therefore, they receive zero AP and are reported separately as an open-set limitation. Long-tail AP_{50} evaluates categories with 6 to 19 training samples, representing low-frequency but still learnable classes. Few-shot AP_{50} evaluates categories with 1 to 5 training samples, representing the most severe closed-set recognition regime.

Head and medium AP_{50} are used as auxiliary diagnostics for frequent-class behavior, whereas the main evidence focuses on overall, class-balanced, long-tail, and few-shot AP_{50}. Recognition-only experiments on ground-truth crops use Top-1, Top-5, and class-balanced accuracy because no bounding-box ranking is involved.

4.4 Main End-to-End Results

All end-to-end comparisons are conducted on the rubbing main-character subset, with seed-42 runs for the full baseline and ablation set and three-seed robustness for the main controlled methods. The central result of this study is not which method scores highest, but what the frequency-aware breakdown in Table 3 reveals: a detector that looks adequate under aggregate AP fails almost completely on rare glyphs. We report the external published reference (Table 2) only to confirm that our baseline is fairly trained and competitive; it is not the object of study.

4.4.1 Baseline Calibration Against the External Reference

The published OBIMD technical validation provides an external reference for the rubbing main-character task [9].

Table 2 Published OBIMD reference and controlled results on the rubbing main-character task.

Method	AP_{50}	mAP@50:95	Prec.	Recall	F1	Avg. IoU
Published OBIMD [9]	0.5141	0.3045	0.6386	0.4271	0.5118	0.7214
Controlled direct YOLO	0.6347	0.3369	**0.8411**	0.5957	0.6975	**0.7835**
FAR-YOLO	**0.6533**	**0.3495**	0.8200	**0.6162**	**0.7036**	0.7799

The published validation reports its first AP column as mAP@50, which our evaluator computes identically as AP_{50} at IoU = 0.50. Our controlled direct YOLO and FAR-YOLO are trained and evaluated with the same data split and evaluation code, so they—not the external row—form the basis for every comparison below. The external reference simply confirms that the controlled baseline operates at or above the performance level of the original validation, ruling out a weak-baseline explanation for the diagnostic effects reported next.

4.4.2 The Central Diagnostic: Aggregate Metrics Conceal Rare-Class Collapse

Table 3 Controlled end-to-end detection and recognition results (seed 42).

Method	Overall AP_{50}	Class-balanced AP_{50}	Long-tail AP_{50}	Few-shot AP_{50}
Direct multi-class YOLO	0.6347	0.3394	0.3340	0.0611
Class-weighted direct YOLO	0.5285	0.2597	0.2317	0.0770
Cross-entropy cascade	0.6458	0.3319	0.2680	0.0790
Prototype-balanced cascade	0.5433	0.3557	0.3583	0.1387
FAR-YOLO	**0.6533**	**0.3767**	**0.3824**	**0.1660**

Read by aggregate AP_{50} alone, the direct multi-class YOLO detector looks adequate, reaching 0.6347 overall AP_{50}. Disaggregated by frequency, the same detector collapses to 0.3340 long-tail and 0.0611 few-shot AP_{50}. **Finding 1:** a single aggregate score hides an almost total failure on exactly the rare glyphs that matter most for paleographic discovery.

Finding 2: the naive in-head remedy backfires. Applying inverse-frequency class reweighting to the dense head—the most direct in-head remedy—lifts few-shot AP_{50} only marginally (0.0611 to 0.0770) while dragging overall AP_{50} from 0.6347 down to 0.5285 and class-balanced AP_{50} from 0.3394 to 0.2597. Forcing the head to emphasize rare classes perturbs localization, objectness, and classification at once, so the global trade-off degrades rather than improves. We test only this naive reweighting; more elaborate long-tailed detection losses, such as equalization loss [34] and seesaw loss [35], may behave differently and are a natural baseline for future work. Our claim is therefore the scoped one—that the simplest in-head fix backfires—which is enough to motivate the decoupled treatment examined next.

Finding 3: decoupling recovers much of the gap but does not close it. Once recognition is separated from localization, the minimal frequency-aware recognizer (FAR-YOLO) raises few-shot AP_{50} from 0.0611 to 0.1660 and class-balanced AP_{50} from 0.3394 to 0.3767 while maintaining overall AP_{50}; a prototype-balanced recognizer improves rare classes but sacrifices overall AP_{50} (0.5433). Decoupling therefore gives the best trade-off among the controlled methods, yet 0.1660 few-shot AP_{50} remains far from usable—the tail is improved, not solved.

4.4.3 Supplemental Conventional Metrics

Table 4 Supplemental conventional detection and recognition metrics.

Method	AP_{50}	AP_{75}	$mAP_{50:95}$	Prec.	Recall	F1	Avg. IoU
Direct multi-class YOLO	0.6347	0.3112	0.3369	**0.8411**	0.5957	0.6975	**0.7835**
Class-weighted direct YOLO	0.5285	0.2615	0.2819	0.7471	0.4971	0.5970	0.7867
Cross-entropy cascade	0.6458	0.3235	0.3458	0.7991	**0.6174**	0.6966	0.7786
Prototype-balanced cascade	0.5433	0.2633	0.2872	0.7452	0.5297	0.6192	0.7798
FAR-YOLO	**0.6533**	**0.3278**	**0.3495**	0.8200	0.6162	**0.7036**	0.7799

The conventional metrics are consistent with the primary AP_{50} comparison. FAR-YOLO obtains the best AP_{50}, AP_{75}, $mAP_{50:95}$, and F1 among the controlled methods, while its recall remains comparable to the cross-entropy cascade. Direct YOLO has higher precision but lower recall, suggesting that the two-stage setting improves candidate coverage at the cost of additional recognition computation. The class-weighted direct detector has lower precision, recall, and F1 than the unweighted detector, which is consistent with the observed degradation caused by detector-head class reweighting.

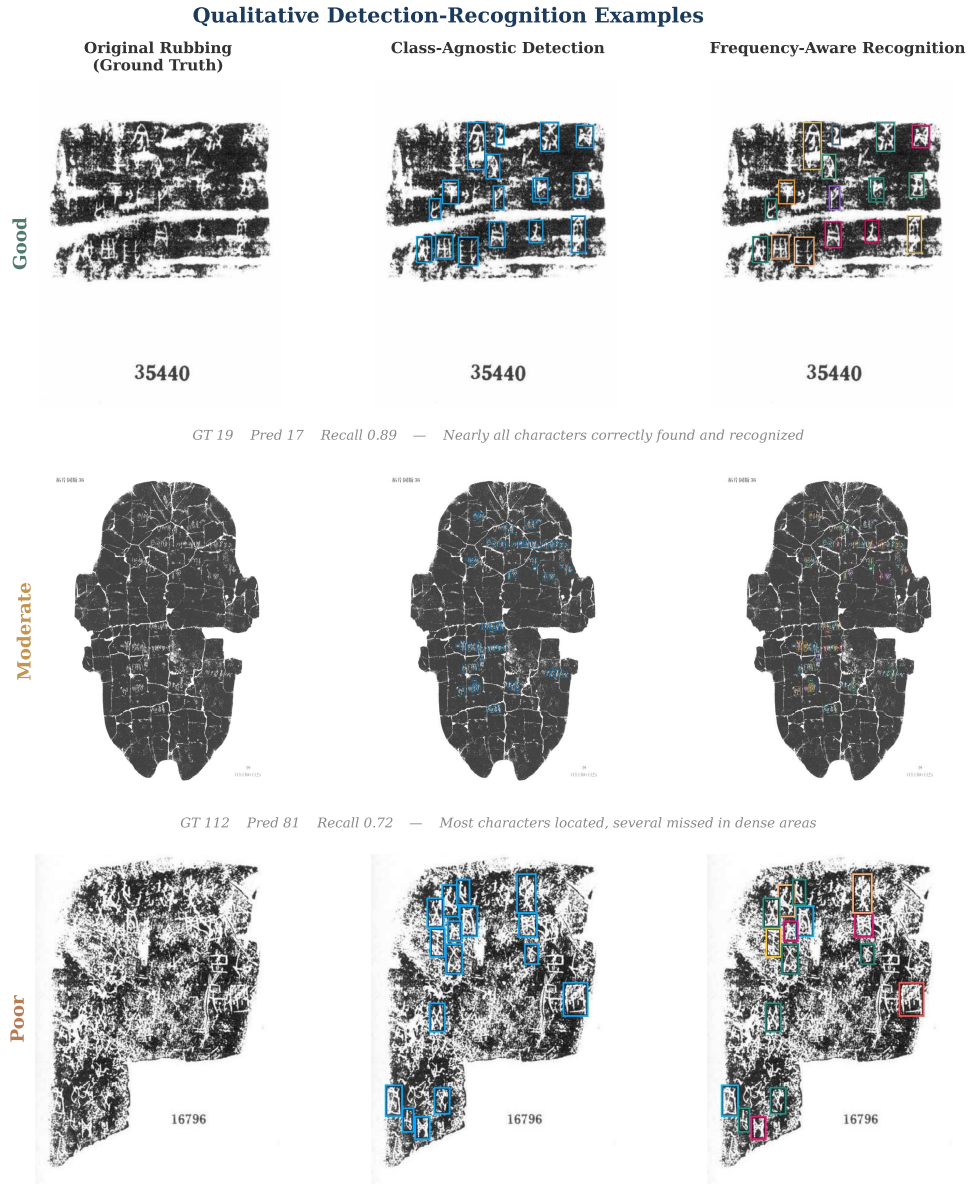


Fig. 4 Qualitative full-page visualization examples on oracle bone rubbings. The figure compares direct multi-class YOLO predictions with FAR-YOLO and highlights cases where rare or few-shot glyphs are corrected.

4.5 Component, Ablation, and Robustness Analysis

4.5.1 Ground-Truth Crop Recognition Results

Table 5 Ground-truth crop recognition results.

Method	Overall Top-1	Class-balanced Acc.	Long-tail Acc.	Few-shot Acc.
Cross-entropy recognizer	0.7692	0.4252	0.3607	0.1019
Prototype-balanced recognizer	0.6686	0.4898	0.4828	0.2427
Augmentation-only recognizer	**0.7842**	0.4501	0.3899	0.0825
Frequency-aware recognizer	0.7773	0.4846	0.4536	0.2136
Nearest-class-mean diagnostic	0.7279	**0.5255**	**0.4907**	**0.2670**

Ground-truth crop recognition isolates the classification component by removing localization errors. Accuracy is used instead of AP because this setting does not involve bounding-box ranking. Prototype-balanced and nearest-class-mean methods produce stronger rare-class results, but they reduce overall Top-1 accuracy. Augmentation alone improves overall Top-1 accuracy, yet it does not improve few-shot recognition. The frequency-aware recognizer improves overall Top-1, class-balanced accuracy, long-tail accuracy, and few-shot accuracy relative to the cross-entropy recognizer, indicating that it provides a more balanced recognition head for the end-to-end cascade.

4.5.2 Separated Detection and Recognition Verification

Localization and ground-truth crop recognition are evaluated separately to determine whether the end-to-end gains originate from the intended components. The class-agnostic YOLO localization branch reaches precision 0.8694, recall 0.7981, AP_{50} 0.8553, and $mAP_{50:95}$ 0.4440 on the same rubbing main-character test setting. This result indicates that visible-character localization is reliable when the detector is not required to predict the long-tailed glyph vocabulary.

The recognition-only comparison further shows that the main bottleneck lies in long-tailed glyph discrimination. Compared with the cross-entropy recognizer, the FAR-YOLO recognizer improves Top-1 accuracy from 0.7692 to 0.7773, class-balanced accuracy from 0.4252 to 0.4846, long-tail accuracy from 0.3607 to 0.4536, and few-shot accuracy from 0.1019 to 0.2136.

4.5.3 Detector-Head Reweighting and Inference Cost

The class-weighted direct YOLO experiment evaluates whether the long-tail problem can be addressed by modifying the YOLO classification loss without decoupling detection and recognition. Its few-shot AP_{50} increases slightly from 0.0611 to 0.0770, but overall AP_{50} decreases from 0.6347 to 0.5285, class-balanced AP_{50} decreases from 0.3394 to 0.2597, and long-tail AP_{50} decreases from 0.3340 to 0.2317. This reinforces

Finding 2: under extreme category imbalance, emphasizing rare classes in the dense detector head damages the global detection-recognition trade-off.

We also measure inference cost on 300 test images using image size 768. Direct multi-class YOLO contains 10.08M parameters and runs at 20.36 ms per image, or 49.12 FPS. FAR-YOLO contains 9.41M detector parameters and 12.03M recognizer parameters, for 21.44M total parameters, and runs at 26.16 ms per image, or 38.22 FPS. FAR-YOLO is therefore slower by about 28.5%, but it remains practical for offline oracle-bone inscription processing.

4.5.4 Ablation Results

Table 6 Ablation results on end-to-end metrics.

Method	**Overall AP_{50}**	**Class-balanced AP_{50}**	**Long-tail AP_{50}**	**Few-shot AP_{50}**
Cross-entropy cascade	0.6458	0.3319	0.2680	0.0790
Oversampling-only cascade	0.6458	0.3520	0.3113	0.1103
Augmentation-only cascade	0.6507	0.3425	0.2882	0.0606
FAR-YOLO	**0.6533**	**0.3767**	**0.3824**	**0.1660**
Logit-adjusted ($\alpha = 0.3$)	0.6390	0.3480	0.3112	0.0986
Logit-adjusted ($\alpha = 0.5$)	0.6278	0.3506	0.3212	0.1021

The ablation results show complementary behavior. Oversampling alone improves class-balanced, long-tail, and few-shot AP_{50} with almost no loss in overall AP_{50}. Augmentation alone improves overall AP_{50}, but it does not improve few-shot AP_{50}. Combining oversampling and morphology-preserving augmentation provides the strongest balance. Logit adjustment improves rare-class metrics relative to the cross-entropy cascade, but stronger adjustment causes a larger overall performance drop. These patterns indicate that the central bottleneck is not only feature extraction or classifier form; it is also the frequency-induced imbalance in training.

The sampling-weight sensitivity analysis tests whether the proposed exposure rebalancing is overly dependent on a single setting.

Table 7 Sampling weight sensitivity.

LT Weight	**FS Weight**	**Overall AP_{50}**	**Class-balanced AP_{50}**	**Long-tail AP_{50}**	**Few-shot AP_{50}**
2	5	0.6508	0.3642	0.3464	0.1321
3	8	**0.6533**	**0.3767**	**0.3824**	**0.1660**
5	10	0.6464	0.3551	0.3599	0.1231

The 3/8 setting gives the best value on all four primary end-to-end metrics. The weaker 2/5 setting preserves overall AP_{50} but provides smaller rare-category gains, whereas the stronger 5/10 setting reduces overall AP_{50} and does not further improve few-shot AP_{50}.

4.5.5 Candidate-Level Diagnostics: The Tail Is "Almost Recognized"

The closed-set AP protocol credits only the top-1 glyph. To test whether rare glyphs are genuinely unrepresented or merely mis-ranked, we re-score the two-stage cascades under top-5 candidate matching, in which a detection counts as correct when the ground-truth glyph appears among the recognizer's five highest-scoring classes (Table 8). The headroom is concentrated almost entirely in the tail: for the cross-entropy cascade, relaxing top-1 to top-5 raises few-shot AP_{50} from 0.0790 to 0.1713 (a 2.2× increase) and long-tail AP_{50} from 0.2680 to 0.4296, while head AP_{50} rises only marginally (0.6617 to 0.7058). The same pattern holds at the recognition level, where overall Top-5 accuracy reaches 0.90 against 0.77 Top-1. Strikingly, simply reading the top-5 candidates of a plain cross-entropy cascade (0.1713 few-shot AP_{50}) recovers as much rare-class performance as the entire frequency-aware training pipeline does at top-1 (0.1660).

Table 8 Top-1 versus top-5 candidate matching for the two-stage cascades. The headroom from allowing ranked candidates is concentrated in the long-tail and few-shot groups, indicating that rare glyphs are often present in the candidate set even when the top-1 label is wrong.

Setting	**Overall AP_{50}**	**Class-balanced AP_{50}**	**Long-tail AP_{50}**	**Few-shot AP_{50}**
Cross-entropy cascade, top-1	0.6458	0.3319	0.2680	0.0790
Cross-entropy cascade, top-5	0.6836	0.4232	0.4296	0.1713
Prototype-balanced cascade, top-1	0.5433	0.3557	0.3583	0.1387
Prototype-balanced cascade, top-5	0.5878	0.4242	0.4257	0.2274

This carries the central practical message of the study: for rare oracle glyphs, the correct identity is frequently present in the model's ranked candidates even when the single top prediction is wrong. The failure of closed-set detection-recognition is therefore, in large part, a failure of forced single-label commitment rather than of visual representation. It directly motivates reformulating the deployed task around candidate ranking, retrieval among visually similar glyphs, and paleographer-in-the-loop verification, rather than around top-1 closed-set accuracy.

4.5.6 Seed Robustness

The three-seed results reduce the risk of drawing conclusions from a single favorable run. Direct multi-class YOLO is stable on overall AP_{50}, but its few-shot AP_{50}

Table 9 Three-seed robustness results (mean ± std).

Method	Overall AP_{50}	Class-balanced AP_{50}	Long-tail AP_{50}	Few-shot AP_{50}
Direct multi-class YOLO	0.6337 ± 0.0009	0.3433 ± 0.0032	0.3373 ± 0.0025	0.0662 ± 0.0043
Cross-entropy cascade	0.6486 ± 0.0022	0.3413 ± 0.0074	0.2926 ± 0.0198	0.0654 ± 0.0097
Prototype-balanced cascade	0.5619 ± 0.0150	0.3640 ± 0.0065	0.3610 ± 0.0025	0.1464 ± 0.0055
FAR-YOLO	**0.6493** ± 0.0029	**0.3709** ± 0.0063	**0.3640** ± 0.0145	**0.1611** ± 0.0196

remains low, confirming that aggregate performance alone does not address rare glyph recognition. Prototype-balanced recognition improves rare-class metrics by changing classifier geometry, but the gain comes with a substantial loss in overall AP_{50}. FAR-YOLO gives the best class-balanced and few-shot averages while keeping overall AP_{50} comparable to the cross-entropy cascade. Compared with the cross-entropy cascade, it improves class-balanced AP_{50} by 0.0296, long-tail AP_{50} by 0.0714, and few-shot AP_{50} by 0.0957 on average. Compared with direct multi-class YOLO, it improves class-balanced AP_{50} by 0.0276, long-tail AP_{50} by 0.0267, and few-shot AP_{50} by 0.0949 on average. These results support the claim that frequency-aware decoupled training improves low-frequency and few-shot categories while largely preserving overall end-to-end performance.

5 Conclusion

This paper diagnosed the task structure of large-vocabulary oracle bone character detection and recognition under the extreme long-tailed distribution of OBIMD. Through a systematic frequency-aware analysis, we arrived at three empirical findings about what makes this task difficult and where the bottleneck actually lies.

First, class-agnostic character localization on rubbing images is already strong ($AP_{50} \approx 0.85$, recall ≈ 0.80)—still imperfect, but far less limiting than recognition—suggesting that, for this task, further detector engineering yields diminishing returns relative to improving rare-glyph recognition. Second, the simplest in-head remedy backfires: applying inverse-frequency class reweighting to the detection head lifts few-shot AP_{50} only marginally while substantially degrading overall and class-balanced performance, though we did not evaluate more sophisticated long-tailed detection losses. Third, a minimal decoupled design—a class-agnostic YOLO localizer paired with a ResNet recognizer trained with few-shot-aware exposure rebalancing—improves few-shot AP_{50} from 0.06 to 0.17 and class-balanced AP_{50} by 3.7 points while maintaining overall AP_{50}. However, even with decoupling and rebalancing, few-shot performance remains far from usable, and candidate-level diagnostics reveal that rare

categories are often present in the top-k predictions even when the top-1 label is incorrect: relaxing top-1 to top-5 matching more than doubles few-shot AP_{50} (0.08 to 0.17) and raises long-tail AP_{50} by 16 points.

These findings carry two practical implications. For evaluation, oracle bone character detection and recognition should be assessed with frequency-aware protocols that disaggregate performance by category frequency, rather than relying solely on aggregate metrics that mask rare-glyph failures. For system design, the task should move beyond closed-set forced classification: practical interpretation pipelines should present ranked candidate lists, integrate retrieval among visually similar glyphs, and support paleographer-in-the-loop verification for low-confidence predictions.

The current study has several limitations. The analysis is confined to the OBIMD rubbing modality and main-character label set—currently the only public resource whose glyph vocabulary is large enough to exhibit this long tail—so cross-dataset validation must await comparably annotated corpora; the gap between 0.17 few-shot AP_{50} and usable accuracy remains large; and the evaluation protocol reports per-group metrics but does not yet offer a single calibrated fairness measure.

Declarations

Funding This research was supported by the National Natural Science Foundation of China (Grant No. 62506007), the Natural Science Foundation of Henan Province (Grant No. 242300420680), the Paleography and Chinese Civilization Inheritance and Development Program (Grant Nos. G1807, G1806, G2821), and the Henan Province Science and Technology Research Project (Grant Nos. 242102210116, 252102321071).
Competing interests The authors declare no competing interests.
Author contributions Conceptualization: D.S., Y.X., and B.L.; methodology: D.S. and Y.X.; software: Y.X.; validation: Y.X.; formal analysis: D.S. and Y.X.; investigation: D.S., Y.X., and P.L.; resources: B.L.; data curation: Y.X.; writing—original draft: D.S.; writing—review and editing: D.S., Y.X., P.L., and B.L.; visualization: P.L.; supervision: B.L.; project administration: B.L.; funding acquisition: B.L. All authors have read and approved the final manuscript.
Ethics approval Not applicable.
Consent to participate Not applicable.
Data availability The OBIMD dataset used in this study is publicly available at https://huggingface.co/datasets/KLOBIP/OBIMD. The processed experiment configurations, evaluation scripts, and result summaries will be made available in a public repository upon acceptance.
Consent for publication Not applicable.
Code availability The source code and experiment configurations will be made available in a public repository upon acceptance.

References

[1] Meng L, Lyu B, Zhang Z, Aravinda CV, Kamitoku N, Yamazaki K (2019) Oracle bone inscription detector based on SSD. In: New Trends in Image Analysis and Processing – ICIAP 2019. Springer, Cham, p 126–136

[2] Xing J, Liu G, Xiong J (2019) Oracle bone inscription detection: A survey of oracle bone inscription detection based on deep learning algorithm. In: Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing. ACM, New York, Article 39, p 1–8. https://doi.org/10.1145/3371425.3371434

[3] Fujikawa Y, Li H, Yue X, Aravinda CV, Prabhu GA, Meng L (2022) Recognition of oracle bone inscriptions by using two deep learning models. Int J Digit Humanit 5:65–79. https://doi.org/10.1007/s42803-022-00044-9

[4] Wen X, Setthawong R, Sun H (2025) Oracle bone inscriptions recognition based on deep learning. In: 2025 IEEE 2nd International Conference on Electronics, Communications and Intelligent Science (ECIS). IEEE. https://doi.org/10.1109/ECIS65594.2025.11086647

[5] Li B, Zhang J, Yu N, Zhang Z, Liu Y, Han Y (2024) Oracle character prototype-guided cyclic disentanglement for oracle bone inscriptions detection. In: Proceedings of ICPRAI 2024

[6] Fu X, Zhou R (2024) Shape prior fusion for oracle bone inscriptions detection. In: Proceedings of the 2024 7th International Conference on Image and Graphics Processing. ACM, New York, p 394–401. https://doi.org/10.1145/3647649.3647711

[7] Fu X, Zhou R, Yang X, Li C (2024) Detecting oracle bone inscriptions via pseudo-category labels. Herit Sci 12:107. https://doi.org/10.1186/s40494-024-01221-5

[8] Li JW, He JR, Wu JR, Liu JM, Lin WB, Huang YS, Wang LJ, Lv JJ, Zeng XX, Chen RJ (2026) Prism-OBI: A novel framework for oracle bone inscription recognition via visual perception and feature decoupling. npj Herit Sci 14:218. https://doi.org/10.1038/s40494-026-02493-9

[9] Li B, Yang J, Liang Y, Hu X, Ding Z, Peng X, Han S, Qin P, Luo D, Jin T, et al (2026) OBIMD: A multi-modal dataset for contextual interpretation of oracle bone inscriptions. Sci Data 13. https://doi.org/10.1038/s41597-026-06967-0

[10] Li J, Wang Q, Huang K, Yang X, Zhang R, Goulermas JY (2021) Mix-Up augmentation for oracle character recognition with imbalanced data distribution. In: International Conference on Document Analysis and Recognition, p 237–251

[11] Li J, Wang Q, Huang K, Yang X, Zhang R, Goulermas JY (2023) Towards better long-tailed oracle character recognition with adversarial data augmentation. Pattern Recognit 140:109534. https://doi.org/10.1016/j.patcog.2023.109534

[12] Li J, Dong B, Wang Q, Huang K, Wang D-H, Liu Y, Liu C-L (2023) Decoupled learning for long-tailed oracle character recognition. In: International Conference on Document Analysis and Recognition, p 165–181

[13] Zhao X, Liu S, Wang Y, Lian Z (2022) FFD augmentor: Towards few-shot oracle character recognition from scratch. In: Asian Conference on Computer Vision, p 37–53

[14] Wang N, Sun Q, Jiao Q, Ma J (2020) Oracle bone inscriptions detection in rubbings based on deep learning. In: 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE

[15] Liu Z, Wang X, Yang C, Liu J, Yao X, Xu Z, Guan Y (2021) Oracle character detection based on improved Faster R-CNN. In: 2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). IEEE

[16] Liu G, Xing J, Xiong J (2020) Spatial pyramid block for oracle bone inscription detection. In: Proceedings of the 2020 9th International Conference on Software and Computer Applications. ACM, New York, p 133–138. https://doi.org/10.1145/3384544.3384561

[17] Zhen Q, Wu L, Liu G (2024) An oracle bone inscriptions detection algorithm based on improved YOLOv8. Algorithms 17(5):174. https://doi.org/10.3390/a17050174

[18] Li Y, Chen H, Zhang W, Sun W (2024) Lightweight oracle bone character detection algorithm based on improved YOLOv7-tiny. In: Proceedings of ICMA 2024

[19] Wang Y, Duan X, Zheng T, Sun Y (2025) Object detection algorithm based on YOLO-PDE. In: Proceedings of CVIDL 2025

[20] Song Y, Zhang C, Wu M (2025) CharDetNet: Lightweight ancient character detection with mixed attention. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN 2025). IEEE

[21] Xiong R, Liu J, Zhang S, Liao Z (2025) FDW-YOLO: An improved YOLOv12 for oracle bone inscriptions detection. In: Neural Information Processing. Springer, Singapore, p 252–267

[22] Gao F, Liu Y, Li D, Chen X, Jiang R, Han Y (2025) Information disentanglement for unsupervised domain adaptive oracle bone inscriptions detection. Signal Process Image Commun

[23] Yue M, Shi D, Diao X, Guo S, Li C, Xu H (2025) Ancient character detection based on fine-grained density map. npj Herit Sci 13:312

[24] Cai J, Li Q, Zhu X, Li B, Yan X, Zhang X (2025) Oracle bone inscription detection with multi-branch feature fusion and efficient attention. In: Proceedings of ATAIT 2025. CEUR Workshop Proceedings 4082, p 150–161

[25] Lin TY, Goyal P, Girshick R, He K, Dollar P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, p 2999–3007. https://doi.org/10.1109/ICCV.2017.324

[26] Kang B, Xie S, Rohrbach M, Yan Z, Gordo A, Feng J, Kalantidis Y (2020) Decoupling representation and classifier for long-tailed recognition. In: International Conference on Learning Representations

[27] Ren J, Yu C, Ma X, Zhao H, Yi S, Li H (2020) Balanced Meta-Softmax for long-tailed visual recognition. In: Advances in Neural Information Processing Systems 33

[28] Menon AK, Jayasumana S, Rawat AS, Jain H, Veit A, Kumar S (2021) Long-tail learning via logit adjustment. In: International Conference on Learning Representations

[29] Li J, Chi X, Wang Q, Huang K, Wang D-H, Liu Y, Liu C-L (2026) A comprehensive survey of oracle character recognition: Challenges, datasets, methodology, and beyond. Pattern Recognit 169:111824. https://doi.org/10.1016/j.patcog.2025.111824

[30] Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D (2016) Matching networks for one shot learning. In: Advances in Neural Information Processing Systems 29

[31] Snell J, Swersky K, Zemel R (2017) Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems 30

[32] Jocher G, Qiu J (2024) Ultralytics YOLO11, version 11.0.0. https://github.com/ultralytics/ultralytics. Accessed 1 Jan 2025

[33] Gupta A, Dollár P, Girshick R (2019) LVIS: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p 5356–5364

[34] Tan J, Wang C, Li B, Li Q, Ouyang W, Yin C, Yan J (2020) Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p 11662–11671

[35] Wang J, Zhang W, Zang Y, Cao Y, Pang J, Gong T, Chen K, Liu Z, Loy CC, Lin D (2021) Seesaw loss for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p 9695–9704