Scientific Data

https://doi.org/10.1038/s41597-026-06967-0

Article in Press

OBIMD: A Multi-modal Dataset for Contextual Interpretation of Oracle Bone Inscriptions

Received: 16 July 2025

Accepted: 24 February 2026

Cite this article as: Li, B., Yang, J., Liang, Y. *et al.* OBIMD: A Multi-modal Dataset for Contextual Interpretation of Oracle Bone Inscriptions. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-06967-0

Bang Li, Jing Yang, Yujie Liang, Xiaobin Hu, Zengmao Ding, Xu Peng, Shengwei Han, Peichao Qin, Donghao Luo, Taisong Jin, Feng Gao, Yongge Liu & Rongrong Ji

SCIENTIFIC DATA

Tracking no: SDATA-25-03932A

OBIMD: A Multi-modal Dataset for Contextual Interpretation of Oracle Bone Inscriptions

Authors: Taisong Jin (Xiamen University), Bang Li (Key Laboratory of Oracle Bone Inscriptions Information Processing, Ministry of Education), Jing Yang (Anyang Normal University), Yujie Liang (Xiamen University), Xiaobin Hu (Tencent), Zengmao Ding (Anyang Normal University), Xu Peng (Tencent), Shengwei Han (Anyang Normal University), Peichao Qin (University of Cambridge), Donghao Luo (Tencent), Feng Gao (Anyang Normal University), Yongge Liu (Key Laboratory of Oracle Bone Inscriptions Information Processing, Ministry of Education), and Rongrong Ji (Xiamen University)

Abstract:
Oracle bone inscriptions, the earliest known form of Chinese writing, hold immense historical and linguistic significance. However, existing digital datasets are typically limited to isolated characters and lack contextual and structural information essential for comprehensive analysis. We present the Oracle Bone Inscriptions Multi-modal Dataset (OBIMD), a large-scale, publicly available corpus to provide pixel-aligned rubbing and facsimile images, character-level annotations, and sentence-level transcriptions with corresponding reading sequences. OBIMD encompasses 10,077 oracle bone inscription images spanning five phases of the Shang Dynasty, featuring 93,652 annotated characters, 21,667 recorded missing-character positions, 21,941 sentence units, and 4,192 non-sentential elements. By integrating visual, structural, and linguistic modalities, OBIMD supports multi-modal learning and diverse tasks such as facsimile enhancement, character retrieval, and syntactic reconstruction. It constitutes a foundational resource for oracle bone inscription recognition and interpretation, enabling scalable and systematic analysis of ancient Chinese writing.

Datasets:

Repository Name	Dataset Title	Accession Number or DOI	URL to data record	Private reviewer access URL/code
Hugging Face Datasets	OBIMD	https://doi.org/10.57967/hf/7821	https://huggingface.co/datasets/KLOBIP/OBIMD	

OBIMD: A Multi-modal Dataset for Contextual Interpretation of Oracle Bone Inscriptions

Bang Li[1,†], Jing Yang[1,†], Yujie Liang[2], Xiaobin Hu[3], Zengmao Ding[1], Xu Peng[3], Shengwei Han[1], Peichao Qin[4], Donghao Luo[3,*], Taisong Jin[2,*], Feng Gao[1], Yongge Liu[1], and Rongrong Ji[2]

[1]Key Laboratory of Oracle Bone Inscriptions Information Processing, Ministry of Education of China, Anyang Normal University
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
[3]Youtu Lab, Tencent
[4]Faculty of Asian and Middle Eastern Studies, University of Cambridge
[*]Corresponding authors: michaelluo@tencent.com, jintaisong@xmu.edu.cn
[†]These authors contributed equally to this work.

Abstract

Oracle bone inscriptions, the earliest known form of Chinese writing, hold immense historical and linguistic significance. However, existing digital datasets are typically limited to isolated characters and lack contextual and structural information essential for comprehensive analysis. We present the Oracle Bone Inscriptions Multi-modal Dataset (OBIMD), a large-scale, publicly available corpus to provide pixel-aligned rubbing and facsimile images, character-level annotations, and sentence-level transcriptions with corresponding reading sequences. OBIMD encompasses 10,077 oracle bone inscription images spanning five phases of the Shang Dynasty, featuring 93,652 annotated characters, 21,667 recorded missing-character positions, 21,941 sentence units, and 4,192 non-sentential elements. By integrating visual, structural, and linguistic modalities, OBIMD supports multi-modal learning and diverse tasks such as facsimile enhancement, character retrieval, and syntactic reconstruction. It constitutes a foundational resource for oracle bone inscription recognition and interpretation, enabling scalable and systematic analysis of ancient Chinese writing.

1 Background & Summary

Oracle bone inscriptions, engraved primarily on tortoise shells and animal bones over 3,600 years ago, represent one of the world's earliest mature writing systems [1]. These inscriptions serve as ritual records of divinatory inquiries and outcomes during the Shang Dynasty [2], preserving invaluable insights into China's ancient royal governance, religious practices, and socio-economic structures. However, oracle bones were intentionally prepared (e.g., drilled with pits) and heat-cracked for divination, rendering excavated remains intrinsically fragile and fragmented. This physical limitation makes routine access to the originals impractical. Consequently, research on oracle bone inscriptions is typically mediated through published compilations rather than direct examination of the artifacts. As illustrated in Fig. 1, a single inscribed oracle bone is commonly documented in these compilations in three core modalities: rubbings, facsimiles, and transcriptions, which scholars consult in combination to establish reading order and interpret the content.

Oracle bone inscription resources have long been published in extensive corpora, yet the underlying evidence is scattered across disparate volumes and editions, making routine lookup, cross-referencing, and multi-modal alignment labor-intensive. To facilitate scholarly access, these materials have been progressively digitized and released through specialized platforms that aggregate and organize records from the published compilations. In 2019, we launched the "Yin Qi Wen Yuan" database [3], which compiles nearly 240,000 oracle bone image records, and links the corresponding rubbings, facsimiles, and transcriptions across publications into unified entries to the extent possible. Nevertheless, despite improved accessibility at the compilation level, the data granularity of digitized oracle bone inscription materials remains critically limited. Rubbings continue to be presented as unsegmented image batches rather than semantically indexed fragments, preventing intelligent algorithms from performing contextualized textual analysis.

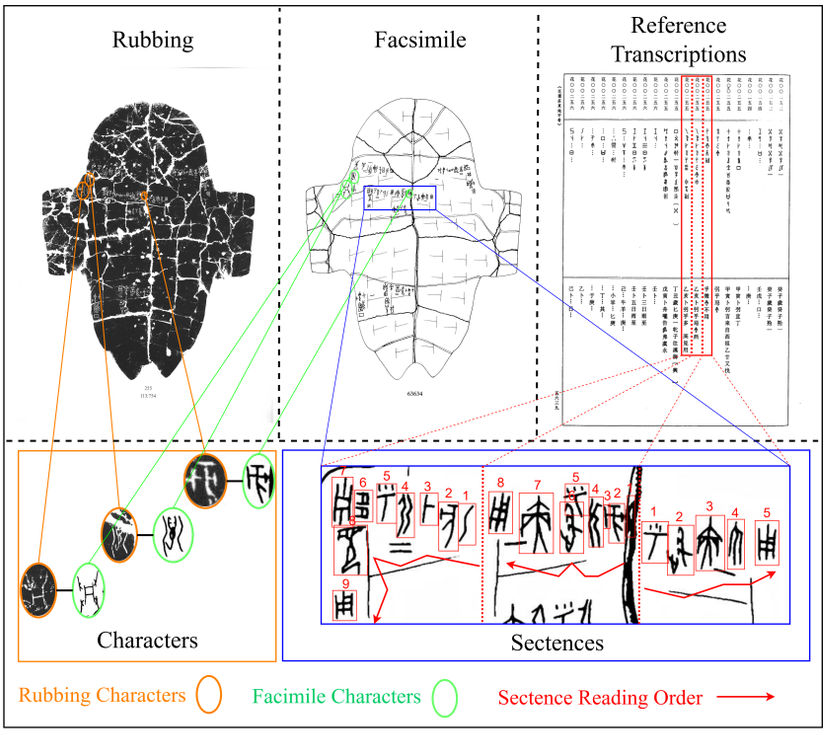


Fig. 1 Three modalities of oracle bone research materials: Rubbing image, Facsimile image, and Reference transcriptions. Magnified views show differences in character appearance between rubbing and facsimile images, while the combination of facsimile and transcription reveals the complex non-linear reading order of oracle bone inscriptions.

Consequently, AI applications are largely restricted to image-based tasks such as character recognition [4,5] or character detection [6,7]. Table 1 highlights that existing oracle bone inscription datasets are predominantly single-modality and built around cropped glyphs, framing the problem as either isolated classification or box-level localization. Rubbing-based recognition datasets (e.g., OBC306 [8], OBI-125 [9], and Oracle-MNIST [10]) typically cover only a limited inventory of well-deciphered characters, whereas facsimile-based datasets (e.g., HWOBC [11], HUST-OBC [12], and EVOBC [13]) rely on expert-rendered, standardized forms that may deviate from the visual evidence in original rubbings. Detection datasets, in turn, often annotate bounding boxes without providing category labels [7].

Dataset	Ann. granularity (level)	Tasks	Year	Full imgs (R/F)	Crops (R/F)	#Class
OBI det. dataset [7]	Char: Box (no cls)	Det.	2020	9,306/0	0/0	N/A
OBC306 [8]	Char: Cls (P)	Rec. (Rub)	2019	0/0	309,551/0	306
OBI-125 [9]			2022	0/0	3,861/0	125
Oracle-MNIST [10]			2024	0/0	30,222/0	10
HWOBC [11]		Rec. (Fac)	2020	0/0	0/83,251	3,881
Oracle-50K [14]			2020	0/0	0/59,081	2,668
Oracle-FS [14]			2020	0/0	0/5,000	200
Ancien-3/5 [15]			2021	0/0	0/39,009	1,186
HUST-OBC [12]			2024	0/0	0/77,064	1,588
EVOBC [13]			2024	0/0	0/75,681	1,762

Dataset	Ann. granularity (level)	Tasks	Year	Full imgs (R/F)	Crops (R/F)	#Cls
Shirakawa Hand-notated OBI [16]	Doc: Crops (cluster; no std labels)	DocOrg	2024	0/1,188	0/370	N/A
OBIMD (Ours) [17]	Multi: Box+PH + Cls (P/S) + Grp + RO	Det.; Rec.; Struct.	2024	10,077/10,077	93,652/93,652	1,730 (P) + 2,747 (S)

Notes. R/F: rubbing/facsimile. Ann. granularity abbreviations: Char=character level; Doc=document level; Box=bounding boxes; Cls=category labels; P/S=primary/sub categories; PH=placeholder (missing position); Grp=group (sentence or non-sentential component); RO=within-group reading order. Task abbreviations: Det.=object detection; Rec.=character recognition; DocOrg=unsupervised document organization. Struct denotes reading-oriented structure annotations, including group assignment (Grp), within-group reading order (RO), and explicit placeholders (PH) for missing positions.

Table 1. Comparison of OBIMD with existing oracle bone inscription datasets in annotation granularity, supported tasks, image modalities, and dataset scale.

However, this crop-centric, single-modality formulation implicitly assumes that a character's identity can be recovered from its local appearance alone. In practical oracle bone inscription interpretation, many glyphs are blurred, eroded, or fragmentary; even specialists may be unable to assign a confident reading from an isolated patch, whereas the same glyph often becomes interpretable once broader evidence is considered. As illustrated in Fig. 2, scholars typically resolve such ambiguity by jointly consulting (i) the surrounding inscription context in the rubbing, (ii) the corresponding facsimile for clearer stroke grounding, and (iii) the transcription as a working reading hypothesis, while also tracking missing or illegible positions that affect sentence completeness. These observations motivate a unified, semantically structured oracle bone inscription dataset—OBIMD [17].

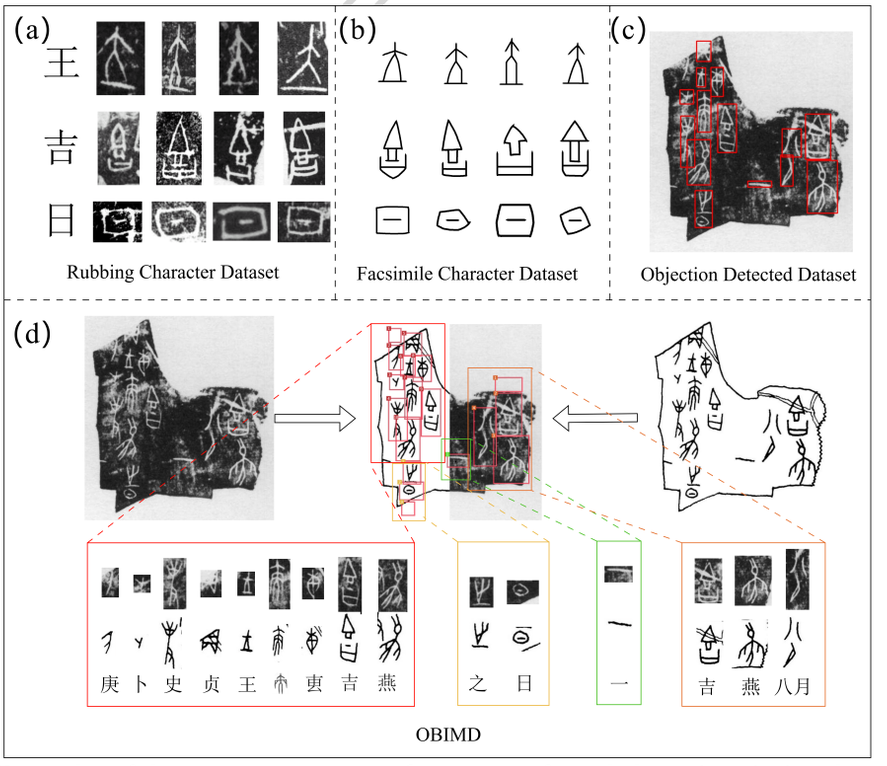


Fig. 2 Examples of data structures across four types of oracle bone inscription datasets. (a) Rubbing character dataset. (b) Facsimile character dataset. (c) Object detection datasets. (d) OBIMD integrates multiple modalities, such as rubbed images, facsimiles, transcriptions, and character, level annotations—while providing sentence groupings and reading order, thereby enabling comprehensive modeling of oracle bone inscription texts.

OBIMD unifies all major oracle bone inscription data modalities into a semantically structured and richly annotated corpus. It includes 10,077 rubbing images from five historical phases of the Shang dynasty, 93,652 annotated characters, 21,667 recorded missing-character positions, 21,941 syntactically validated sentences, and 4,192 non-sentential groups. These data are aligned across rubbings, facsimiles, and transcriptions using a dedicated annotation platform [18] equipped with reference tools and verified by domain experts. OBIMD thus fills a critical gap in existing resources and enables comprehensive modeling of oracle bone texts.

Constructing such a dataset presents substantial challenges. Oracle bone inscriptions are often fragmentary, visually ambiguous, and syntactically nonlinear, requiring deep paleographic expertise to interpret. However, qualified experts are few, and large-scale manual annotation is prohibitively time-consuming. To address this, we developed a scalable collaborative pipeline that integrates annotators with varying levels of expertise, supported by cross-referencing tools and expert validation mechanisms. This framework balances quality with scalability, and ensures that OBIMD serves not only as a reliable research resource, but also as a foundation for AI systems that model oracle bone texts beyond isolated glyphs. In particular, the inclusion of sentence-group annotations with reading order and explicit placeholders for missing positions enables learning and evaluation of structure-aware interpretation under fragmentation.

2 Methods

To construct a multi-modal oracle bone inscription dataset while addressing expert scarcity and annotation complexity, we designed an efficient collaborative annotation platform[18] that supports the entire pipeline. As depicted in Fig. 3, our framework follows a three-stage collaborative workflow: it begins with data acquisition to aggregate rubbings, facsimiles, and reference transcriptions with structured indexing capabilities. This is followed by pre-annotation using domain-specific datasets to generate preliminary character annotations. The collaborative annotation and verification phase enables non-specialists to refine labels via a transcription-assisted platform that supports fragmented glyph queries and cross-modal comparisons. Graduate annotators then conduct cross-reference validation, while expert arbitration is reserved exclusively for annotations that remain unresolved through iterative, reference-informed graduate review.

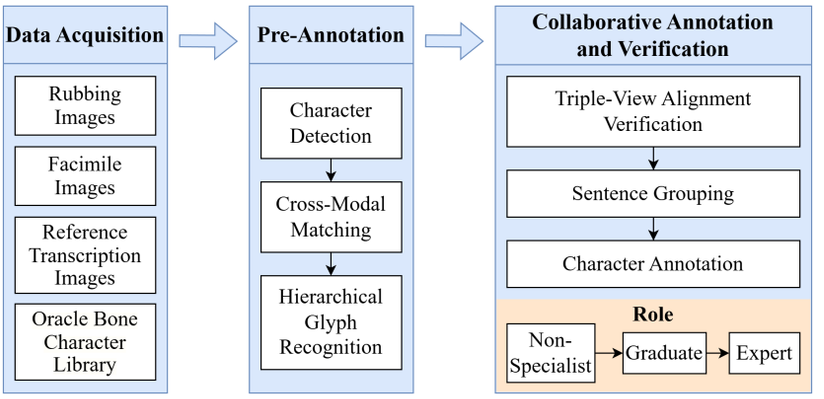


Fig. 3 Pipeline for constructing the OBIMD dataset.

Data acquisition.

Rubbings, serving as the foundational multi-modal data for oracle bone inscription studies, were systematically acquired from the "Yin Qi Wen Yuan" digital repository [3]. The dataset contains 10,077 curated rubbing images, with 9,913 specimens sourced from "Collection of Oracle Bone Inscriptions" [19] the most comprehensive archival compendium of such materials. Given the inherent fragmentation of original oracle bones, which predominantly yields discontinuous rubbing samples, we strategically augmented the dataset with 164 high-integrity rubbings derived from the Oracle Inscriptions from "Yinxu Huayuanzhuang East Oracle Bones" [20].

The facsimile serves not only as the researchers' interpretation of oracle bone information but also as an essential medium for non-specialists to identify annotation targets. To ensure the accuracy and authority of the annotations, non-specialists refer to the facsimiles corresponding to the rubbings in the "Comprehensive Series of Oracle Bone Facsimiles" [21]. However, the pixel-aligned facsimiles in the dataset do not originate directly from this series. This discrepancy arises because the facsimiles in the Comprehensive Series contain non-textual patterns that are not evident in the rubbings, as illustrated in Fig. 4. To ensure consistency between the rubbings and the facsimiles, the facsimile

images in the dataset have been redrawn by integrating the selected rubbings with the facsimiles from the Comprehensive Series.

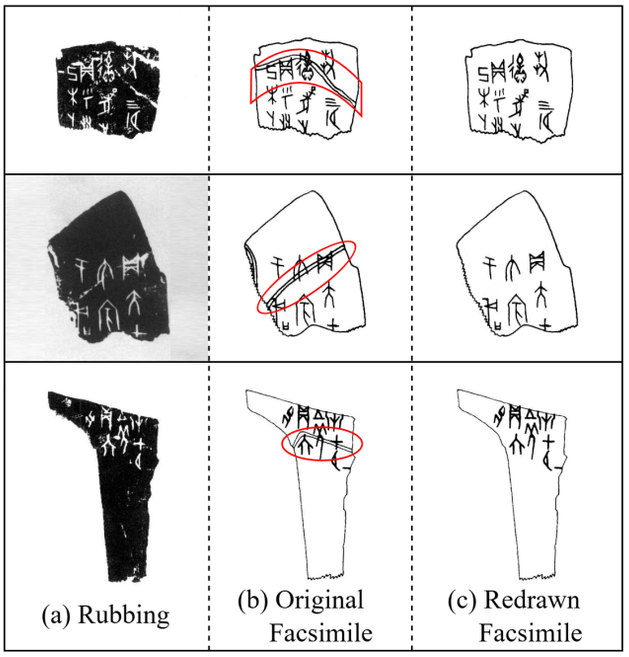


Fig. 4 Comparison of rubbing, published facsimile, and our redrawn version. Non-character patterns like shield-shaped grooves, unclear in the rubbing but emphasized in the original facsimile, are removed in our redrawings to avoid misrepresenting characters.

In addition to the core dataset materials, the annotation platform provides scanned pages from transcription volumes available in the "Yin Qi Wen Yuan" repository [3] as references for annotating the aligned rubbing–facsimile pairs. These pages present inscription glyphs together with their transcriptions in a conventional sentence-level reading order, enabling annotators to locate corresponding regions on the rubbings/facsimiles and reconstruct reading sequences from direct visual evidence. Because a single oracle bone fragment may contain multiple independent inscription units, annotators first assign characters to their corresponding groups (primarily sentential groups, but also non-sentential components when applicable) and then annotate the within-group reading order. Moreover, fragmentation often leaves sentences incomplete; such missing positions are recorded as placeholders (empty boxes placed near breaks) to indicate where a character is expected in the reading sequence but not visible.

For character-category annotation, the platform integrates the Jingyuan oracle bone character library from the "Oracular Digital Platform" [22]. This library adopts a two-tier hierarchy in which primary headings define main character entries and secondary headings distinguish major form variants as sub-characters that may differ markedly in component configuration or stroke realization. Annotators are not required to explicitly reason about the hierarchy; instead, they select the closest-matching entry based on visual appearance and the aligned references, and the main–sub relationship is recorded according to the library taxonomy to retain fine-grained distinctions that should not be collapsed into a single modern-character label. Overall, this design yields a three-level hierarchy from images to characters and then to sentential groups.

Pre-annotation. To significantly improve annotation efficiency by enabling non-expert annotators to quickly locate and identify characters, we designed a three-stage automated pre-annotation pipeline for paired rubbing and facsimile images. This pipeline reduces the initial annotation burden by providing accurate candidate regions and labels, thus accelerating the overall annotation process while maintaining data quality.

First, character regions are independently detected on both rubbing and facsimile images using YOLO-based object detectors [23], resulting in two sets of bounding boxes for each image pair.

Second, a cross-modal box matching module aligns detected regions between rubbing and facsimile images. This process constructs a cost matrix that integrates normalized spatial distances and visual similarities of glyph features, and determines the optimal correspondence via the Hungarian algorithm [24]. Cases with mismatches or low-confidence alignments are flagged for manual verification to ensure precise positional alignment.

Third, glyphs within the matched bounding boxes are recognized by matching their shapes to entries in the Oracular Digital Platform character library. For each glyph, the top-k candidate labels and associated sub-labels—based on matched character components—are generated to maintain the hierarchical classification structure.

The pipeline produces structured annotations comprising bounding box coordinates, top candidate labels with confidence scores, and aligned rubbing-facsimile image pairs. These pre-annotated outputs are integrated into a collaborative annotation platform, enabling non-specialist annotators to efficiently review and refine labels with a reduced cognitive burden relative to manual annotation from scratch.

Collaborative annotation and verification. An accessible web-based annotation platform[18] has been developed to support collaborative verification workflows across expertise levels, from non-specialists to domain experts. The platform integrates synchronized image visualization and query functionalities, enabling efficient annotation of bounding box coordinates, attributes, and classifications. Through real-time comparison of rubbing-facsimile pairs and dynamic transcription referencing, non-specialists have demonstrated a 60-fold productivity increase over manual annotation methods.

(a) Triple-View Alignment Verification (b) Sentence Grouping (c) Character Annotation

Fig. 5 Interface of the collaborative annotation platform: (a) Tri-modal interface enabling pixel-level comparison between rubbing and facsimile, with the reference transcription displayed on the side for alignment verification. (b) Sentence grouping interface allowing classification and ordering of characters with color-coded boxes and interactive controls. (c) Character annotation interface integrating a top-10 retrieval algorithm combining oracle bone images and modern transcriptions, plus options for search by character or handwriting input.

The platform's image visualization interface is shown in Fig. 5(a). A dynamic overlay for pixel-level comparison between an oracle bone rubbing and its corresponding facsimile. Meanwhile, the corresponding scanned page of transcription can be displayed in a split-view layout. The annotation interface for the adjustment and attribution of the bounding box is depicted in Fig. 5(b). The bounding boxes generated by pre-annotation can be manually added, deleted, and adjusted in the interactive graphical interface. Meanwhile, these boxes can be grouped into a sentence or a non-sentential component, and reordered via drag-and-drop within each group to match the intended reading sequence as reflected in the reference transcriptions.

The character annotation of the platform is illustrated in Fig. 5(c). The platform supports character category annotation by clicking directly on individual character images. These clickable images are drawn from the top 10 shape similarity matches retrieved from the character database, based on either character screenshots or user-drawn input. This image-based interaction not only improves annotation efficiency and usability but is also necessary, as most undeciphered oracle bone characters lack Unicode encodings, making traditional text-based annotation methods unfeasible. The platform also allows annotators to manually place bounding boxes in regions corresponding to missing characters and label them as placeholders. Additionally, the platform supports the marking of special symbols to flag cases where annotators are unable to find a suitable category, thereby enabling expert review.

To support collaborative annotation and verification, the platform defines three user roles: non-specialist, graduate, and expert. Each user level has access to the annotations submitted by the preceding level. Non-specialist annotators work with data that have been pre-annotated by the algorithm, using reference materials and tools provided by the platform to further complete the manual annotation. When information cannot be reliably resolved using the available references, they assign numeric identifiers to indicate uncertainty. Graduate annotators, who have received formal training in oracle bone studies and are able to read rubbings, are responsible for reviewing the annotations made by non-specialists and resolving uncertain cases. Expert annotators then

review the annotations validated by graduates to ensure overall accuracy and correct any remaining errors.

3 Data Records

The dataset is available at HuggingFace (https://huggingface.co/datasets/KLOBIP/OBIMD [17]), which includes images, JSON files, and tabular data. As shown in Fig. 6, the annotation structure follows a three-level hierarchy: image-level, sentence-level, and character-level. Definitions of all annotated fields are provided in Table 2. The JSON annotation file contains 10,077 oracle bone entries. Each entry includes the catalog abbreviation of the rubbing, as well as all identified groups within the fragment, including both sentence-level units and non-sentential components.

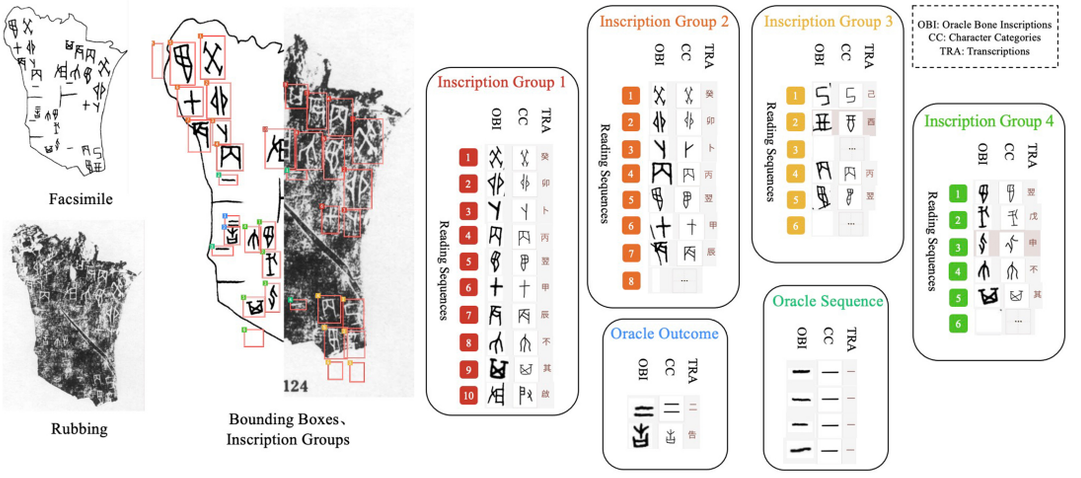


Fig. 6 Hierarchical annotation structure of the OBIMD dataset, covering image-level sources, character-level bounding boxes and annotations, sentence-level groupings, and reading sequences.

Within each entry, character-level bounding boxes are grouped into sentence-like units. Each group is assigned a GroupCategory attribute indicating its type. Sentence groups used for divination are labeled using the format InscriptionSentence1, InscriptionSentence2, etc., where the numeric suffix reflects a conventional reading order among sentence groups. Other group types at the same hierarchical level include non-sentential components: OracleSequence (characters representing the order of divination), OracleOutcome (characters indicating divinatory results), and Uncertain (characters isolated due to fragmentation that cannot be assigned to any semantically defined group). Each oracle bone fragment contains between 1 and 63 sentences or component groups. In total, the dataset comprises 21,941 syntactically validated InscriptionSentences, along with 3,785 OracleSequences and 407 OracleOutcomes.

Each sentence group contains several fields: the “Position” field specifies each box’s location and size in the format (x, y, width, height), while the “OrderNumber” field indicates its reading order within the sentence. Because oracle bone characters contain a wide variety of character variants, the annotation adopts a two-level structure with a main Label and a SubLabel. The SubLabel is used to distinguish different written forms of the same character. This two-level structure is based on the character system used by the Oracular Digital Platform [22], and corresponds directly to the Label and SubLabel fields in our annotation schema.

It should be noted that some bounding boxes are not assigned character labels. For instance, when a character is missing from a sentence due to fragmentation, the corresponding bounding box is marked as a placeholder and given a “SeatFont” set to 1 (default is 0 for non-placeholders). Apart from placeholders, unlabeled boxes carry a “Mark” field to indicate four types of exceptions: 0 for characters too damaged to be identified even by experts; 1 for disputed characters without consensus on classification; 2 for those visible only in the rubbing; and 3 for those found only in the facsimile (with -1 indicating a regular character by default). Although most of these cases correspond to blank or fragmented regions, expert annotators may still assign character labels when contextual evidence supports it. The dataset contains a total of 115,319 bounding boxes, including 93,652 characters and 21,667 placeholders. This includes the detection boxes, character categories, transcriptions, corresponding inscription groups, and reading sequences in the groups of each oracle bone character, contentious or missing parts.

Additional files are provided to supplement the character category information. These include (i) images of sub-characters, (ii) two mapping tables—one linking each sub-character to its corresponding main category and the other mapping each sub-character to the platform-specific glyph code point used by the font of the Oracular Digital Platform[22]—and (iii) a JSON file that records, for each main character, a platform-provided reference modern Chinese character for practical lookup. It should be noted that this modern Chinese character transcription reflects the current state of interpretation; it is provided for reference and convenient lookup only, and should not be treated as final, because oracle-bone decipherment remains ongoing. OBIMD adopts the unique identifiers (UIDs) defined in the Jingyuan oracle bone character library to index both main characters and sub-characters consistently across the mapping tables, OBIMD annotation JSON files and folder structure. Accordingly, sub-character images are stored in a two-level directory structure: the top-level folder is named by the UID of the main character, and each subfolder is named by the UID of the corresponding sub-character. These resources clarify the structural relationships between different character forms and support downstream tasks such as character matching and generation.

Field Name	Description	Example
`Facsimile`	Path to the oracle bone facsimile image	`/facsimile/h00002.jpg`
`Rubbing`	Path to the ink rubbing image	`/rubbing/h00002.jpg`
`RubbingName`	Short identifier for the rubbing image	`H2`
`GroupCategory`	Sentence type or grouping category	`InscriptionSentence1`
`Position`	Bounding box in format (x, y, width, height)	`558,581,80,218`
`OrderNumber`	Order of the character in the sentence	`5`
`Label`	Main character label (used for classification)	`xkubtjk815`
`SubLabel`	Secondary label (often same as Label)	`xkubtjk815`
`SeatFont`	Font type indicator (0: normal, 1: variant)	`0`
`Mark`	Special marker (e.g., -1 means unmarked)	`-1`

Table 2. Data structure of a single case entry in the annotated Multi Modal Oracle Dataset.

4 Technical Validation

To validate the reliability of OBIMD in advancing AI understanding of oracle bone inscriptions, we decompose the expert reading process of oracle bone rubbings into three subtasks: character detection and recognition, layout-based character clustering, and sentence-level character reordering. For each subtask, we adopt a supervised learning paradigm and demonstrate that state-of-the-art algorithms can benefit significantly from our dataset. Meanwhile, the performance gap on more complex cases, such as character subcategories and non-linear layouts, highlights the inherent challenges of OBI understanding and suggests valuable directions for future research.

To assess the applicability of OBIMD for training AI models on oracle bone inscriptions, we design a benchmark task for character-level detection and recognition. This task requires jointly localizing and classifying characters across full oracle bone images. To reflect practical scenarios and maximize coverage of character types, we define four subtasks based on image modality (rubbing vs. facsimile) and character type (main vs. sub-character). Specifically, these subtasks involve recognizing main characters and sub-characters in both rubbings and facsimiles. The dataset is split into training and validation sets in a 9:1 ratio by oracle bone ID, ensuring non-overlapping instances across splits. Bounding boxes marked as placeholders (with "SeatFont" = 1) are excluded from training and evaluation. Each subtask is trained independently using YOLOv11l, a state-of-the-art model for object detection and classification. Model performance is evaluated using standard metrics, including mAP@50, mAP@50:95, precision, recall, F1 score, and average IoU. All results are reported at the 200th training epoch.

As shown in Table 3, OBIMD enables high-performance character detection and recognition across all subtasks, confirming its suitability as a training resource. Among the four subtasks, facsimile main characters yield the best performance (mAP@50 = 0.6666, F1 = 0.6395), while rubbing sub-characters pose the greatest challenge (mAP@50 = 0.4284, F1 = 0.4493). These results reflect domain-specific characteristics of the dataset: facsimile images tend to exhibit clearer glyph contours than rubbings, and main characters are more structurally stable than their sub-character variants. The performance gap between these settings highlights the intrinsic difficulty of oracle bone understanding, especially in degraded rubbing images and for complex or irregular sub-character forms. These findings suggest that OBIMD not only supports model training but also provides challenging benchmarks for future research on ancient script recognition.

To further assess the applicability of OBIMD in syntax-level understanding tasks, we formulate a sentence-level character clustering task that aims to group characters belonging to the same sentence based on their categorical labels and spatial positions. This task is framed as a supervised

Subtask	mAP@50	mAP@50:95	Precision	Recall	F1 Score	Average IoU[1]
Rubbing (Main Characters)	0.5141	0.3045	0.6386	0.4271	0.5118	0.7214
Rubbing (Sub-characters)	0.4284	0.2563	0.5627	0.3739	0.4493	0.7118
Facsimile (Main Characters)	0.6700	0.4470	0.6473	0.6134	0.6395	0.7560
Facsimile (Sub-characters)	0.5853	0.3904	0.6148	0.5245	0.5630	0.7564

[1] Computed by averaging the maximum IoU of each predicted bounding box with all ground truth boxes.

Table 3. Character detection and recognition results on OBIMD.

clustering problem and evaluated using four standard metrics: Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Purity.

Experiments are conducted on the main-character subset and the sub-character subset, with model outputs compared against expert-annotated sentence groupings. As shown in Table 4, the model achieves a high Purity score of 0.84, indicating strong intra-cluster consistency—most characters grouped together indeed belong to the same ground-truth sentence. However, the relatively lower AMI (0.54) and ARI (0.53) scores suggest that sentence boundaries are not always accurately recovered, with some inter-sentence confusion remaining. The NMI score of 0.60 further reflects moderate agreement between predicted and gold clusters, highlighting that the task remains challenging, especially when sentence structures are densely arranged or spatially overlapping. These results underscore both the utility of OBIMD for sentence-level modeling and the open research challenges in the structural analysis of oracle bone texts.

Subtask	AMI	NMI	ARI	Purity
Main Characters	0.54	0.6	0.53	0.84
Sub-characters	0.51	0.53	0.50	0.79

Table 4. Sentence-level character clustering results on OBIMD.

To evaluate the capacity of OBIMD in supporting sentence-level sequential understanding, we design a character reordering task that aims to recover the original reading order of characters within a sentence. Unlike conventional language modeling tasks that leverage standardized character encodings such as Unicode, oracle bone script contains a large number of undeciphered or structurally ambiguous characters, making existing pretrained language models inapplicable. To address this, we formulate the task as a position classification problem: given a shuffled sentence of length ≥ 3, the model predicts the original position index for each character.

We adopt a Transformer-based model with a final linear layer to predict positional indices. The dataset is split into training and validation sets using the same oracle bone ID-based strategy as in the previous tasks to ensure consistency. Performance is evaluated using top-1 accuracy, top-3 accuracy, and Average Distance Error (ADE), which measures the average deviation between predicted and ground-truth positions.

As shown in Table 5, the model achieves strong performance on both main and sub-character subsets. On main characters, it attains a top-1 accuracy of 75.35%, a top-3 accuracy of 91.57%, and an ADE of 0.4974. Sub-characters yield comparable results with a top-1 accuracy of 72.78%, a top-3 accuracy of 90.35%, and an ADE of 0.5260. These results demonstrate the effectiveness of OBIMD in training models for character-level reordering, while the remaining prediction error indicates ongoing challenges in modeling non-linear reading paths and glyph ambiguity.

Subtask	top-1 accuracy	top-3 accuracy	ADE
Main Characters	75.35%	91.57%	0.4974
Sub-characters	72.78%	90.35%	0.5260

Table 5. Sentence-level character reordering results on OBIMD.

5 Data Availability

The OBIMD dataset generated and analysed during the current study is available on the Hugging Face Hub [17].

6 Code Availability

Source code and scripts used for the technical validation experiments on the OBIMD dataset[17] are publicly available on GitHub at https://github.com/libang1991/OBIMD. The repository includes the core implementations of the baseline models evaluated in this manuscript, supporting reproducibility of the reported results. The web-based annotation platform used for OBIMD construction is available at https://www.jgwlbq.org.cn/oracle-bone.

7 References

1. Boltz, W. G. Early Chinese writing. *World Archaeology* **17**, 420–436 (1986).

2. Keightley, D. N. The Shang state as seen in the oracle-bone inscriptions. *Early China* **5**, 25–34 (1979) .https://doi.org/10.1017/S0362502800006118

3. Yin Qi Wen Yuan. Yin Qi Wen Yuan (Oracle Bones Corpus) .https://jgw.aynu.edu.cn/ (2019).

4. Fujikawa, Y., Li, H., Yue, X. *et al.* Recognition of oracle bone inscriptions by using two deep learning models. *International Journal of Digital Humanities* **5**, 65-79 (2023) .https://doi.org/10.1007/s42803-022-00044-9

5. Li, J., Wang, Q. F., Huang, K. *et al.* Towards better long-tailed oracle character recognition with adversarial data augmentation. *Pattern Recognition* **140**, 109534 (2023) .https://doi.org/10.1016/j.patcog.2023.10953

6. Fu, X. & Zhou, R. Shape prior fusion for oracle bone inscriptions detection. in *Proceedings of the 2024 7th International Conference on Image and Graphics Processing* 394-401 (2024) .https://doi.org/10.1145/3647649.3647711

7. Liu, G., Xing, J. & Xiong, J. Spatial pyramid block for oracle bone inscription detection. in *Proceedings of the 2020 9th International Conference on Software and Computer Applications* 133-140 (2020) .https://doi.org/10.1145/3384544.3384561

8. Huang, S., Wang, H., Liu, Y. *et al.* OBC306: A large-scale oracle bone character recognition dataset. in *2019 International Conference on Document Analysis and Recognition (ICDAR)* 681-688 (IEEE, 2019) .https://doi.org/10.1109/ICDAR.2019.00114

9. Yue, X., Li, H., Fujikawa, Y. *et al.* Dynamic dataset augmentation for deep learning-based oracle bone inscriptions recognition. *ACM Journal on Computing and Cultural Heritage* **15**, 1-20 (2022) .https://doi.org/10.1145/3532868

10. Wang, M. & Deng, W. A dataset of oracle characters for benchmarking machine learning algorithms. *Scientific Data* **11**, 87 (2024) .https://doi.org/10.1038/s41597-024-02933-w

11. Li, B., Dai, Q., Gao, F. *et al.* Hwobc-a handwriting oracle bone character recognition database. *Journal of Physics: Conference Series* **1651**, 012050 (IOP Publishing, 2020) .https://doi.org/10.1088/1742-6596/1651/1/012050

12. Wang, P., Zhang, K., Wang, X. *et al.* An open dataset for oracle bone character recognition and decipherment. *Scientific Data* **11**, 976 (2024) .https://doi.org/10.1038/s41597-024-03807-x

13. Guan, H., Wan, J., Liu, Y. *et al.* An open dataset for the evolution of oracle bone characters: Evobc. Preprint at https://arxiv.org/abs/2401.12467 (2024).

14. Han, W., Ren, X., Lin, H. *et al.* Self-supervised learning of orc-bert augmentator for recognizing few-shot oracle characters. in *Proceedings of the Asian Conference on Computer Vision* (2020) .https://doi.org/10.1007/978-3-030-69544-6_39

15. Zhang, G., Liu, D., Smyth, B. *et al.* Deciphering ancient Chinese oracle bone inscriptions using case-based reasoning. in *International Conference on Case-Based Reasoning* 309-324 (Springer, 2021) .https://doi.org/10.1007/978-3-030-86957-1_21

16. Yue, X., Wang, Z., Ishibashi, R. *et al.* An unsupervised automatic organization method for Professor Shirakawa's hand-notated documents of oracle bone inscriptions. *International Journal on Document Analysis and Recognition (IJDAR)* **27**, 583-601 (2024) .https://doi.org/10.1007/s10032-024-00463-0

17. Key Laboratory of Oracle Bone Inscriptions Information Processing, Li, B., Yang, J. *et al.* OBIMD. *Hugging Face* .https://doi.org/10.57967/hf/7828 (2026).

18. Oracle Bone AI Collaborative Platform. Oracle bone AI collaborative platform .https://www.jgwlbq.org.cn/oracle-bone (2024).

19. Guo, M. & Hu, H. (eds) *Jiaguwen Heji* [Collection of Oracle Bone Inscriptions] (Zhonghua Book Company, 1978–1982).

20. Chinese Academy of Social Sciences. *Yinxu Huayuanzhuang East Oracle Bones* (Yunnan Nationalities Publishing House, 2003).

21. Huang, T. *Comprehensive Series of Oracle Bone Facsimiles* (Peking University Press, 2022).

22. Oracular Digital Platform. Glyph library .https://oracular.azurewebsites.net/glyphs (2024).

23. Ultralytics. ultralytics (v8.2.94). *GitHub* .https://github.com/ultralytics/ultralytics/releases/tag/v8.2.94 (2024).

24. Kuhn, H. W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2**, 83–97 (1955) .https://doi.org/10.1002/nav.3800020109

8 Author Contributions

Bang Li and Jing Yang co-wrote the manuscript and contributed equally to this work. Yujie Liang and Zengmao Ding also contributed to manuscript writing. Technical validation experiments were conducted by Bang Li, Jing Yang, and Zengmao Ding. Xiaobin Hu and Taisong Jin proposed key revisions before submission. Bang Li and Donghao Luo initiated and supervised the construction of the OBIMD dataset. Yujie Liang, Zengmao Ding, and Xu Peng implemented algorithmic pre-annotation for the dataset. Bang Li, Jing Yang, and Shengwei Han conducted manual annotation and developed annotation guidelines. Bang Li, Donghao Luo, and Yongge Liu coordinated the annotation teams. Peichao Qin provided the standard oracle bone character library for the dataset. Rongrong Ji, Feng Gao, and Yongge Liu supported the project through funding and resources. Correspondence should be addressed to Donghao Luo or Taisong Jin, who provided overall guidance on project design, dataset framework, and manuscript review. All authors reviewed and approved the final manuscript.

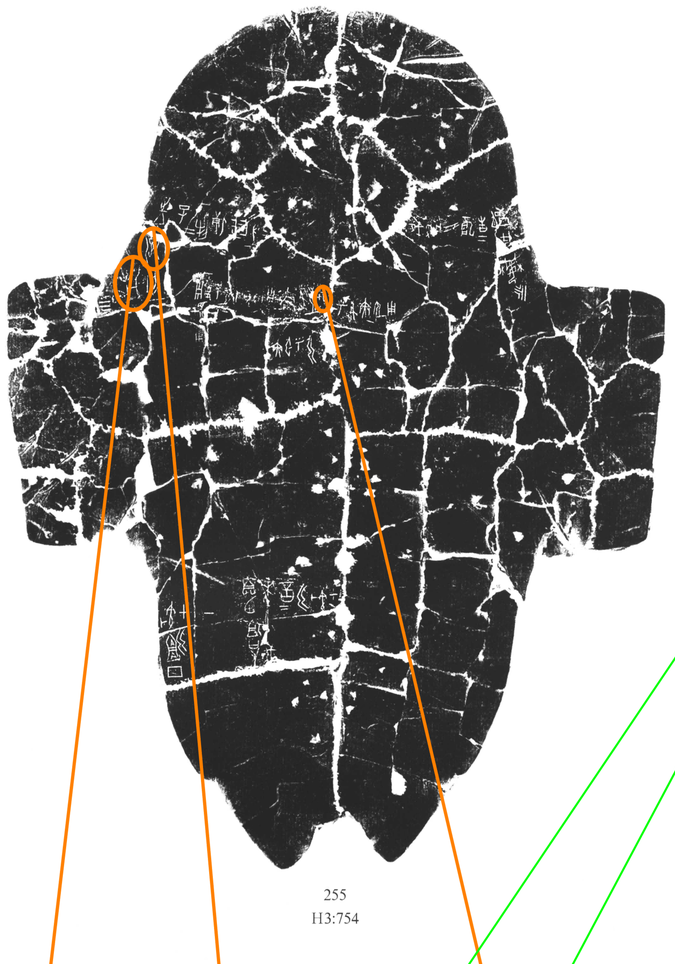
9 Competing Interests

The authors declare no competing interests.

10 Funding

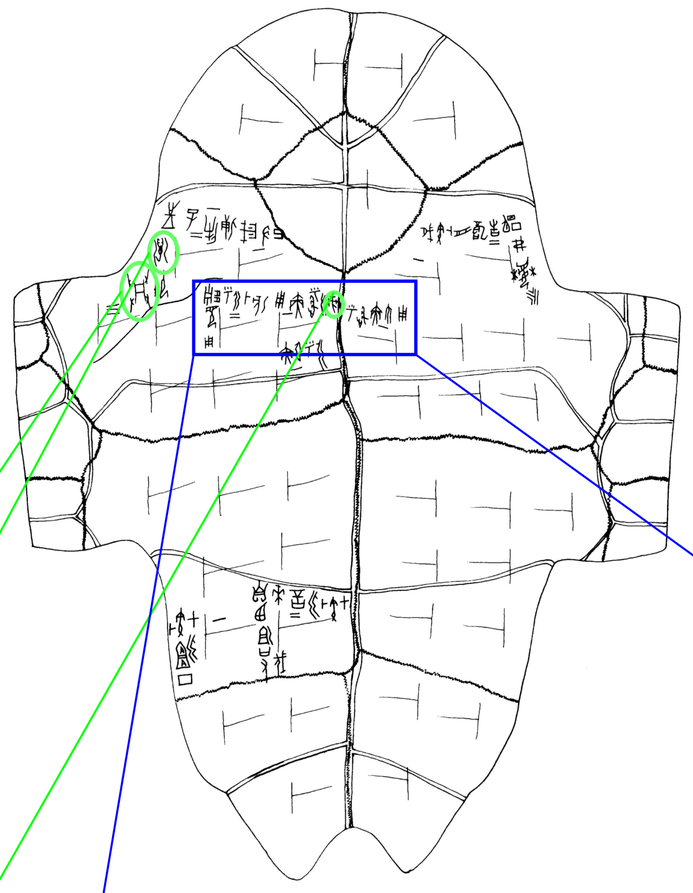
This research was supported by the National Natural Science Foundation of China (Grant No. 62506007), the Natural Science Foundation of Henan Province (Grant No. 242300420680), the Paleography and Chinese Civilization Inheritance and Development Program (Grant Nos. G1807, G1806, G2821), the Henan Province Science and Technology Research Project (Grant Nos. 242102210116, 252102321071), the Open Research Topic of the Key Laboratory of Oracle Information Processing, Ministry of Education (Grant No. OIP2024E002, OIP2024H002), the Key Technology Project of Henan Educational Department of China (Grant No. 22ZX010), and the Henan Province High-Level Talents International Training Program (Grant No. GCC2025028).

Rubbing



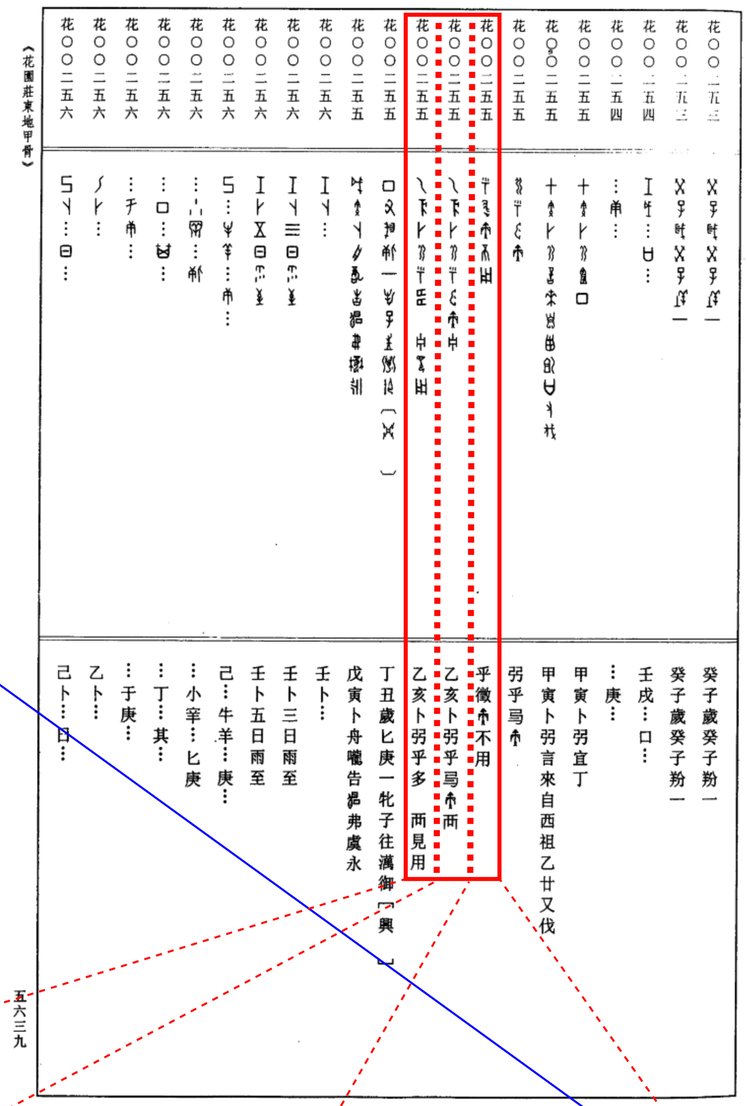
255
H3:754

Facsimile



63634

Reference Transcriptions

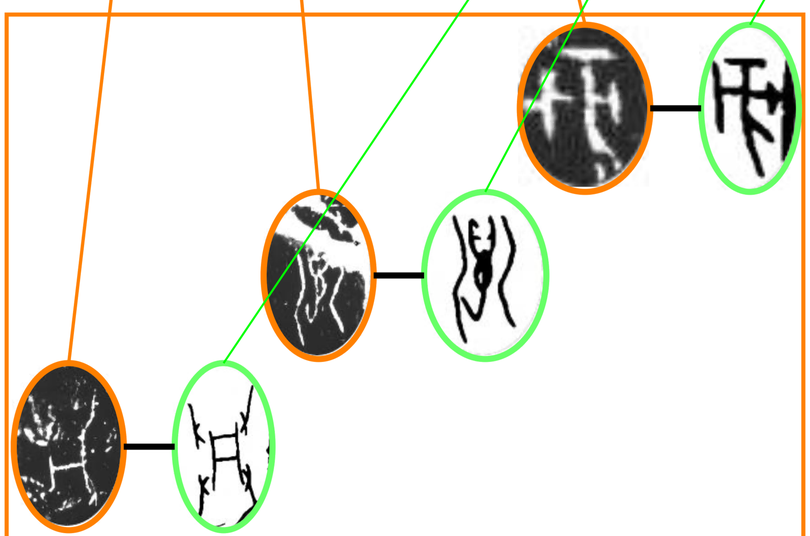


《花園莊東地甲骨》

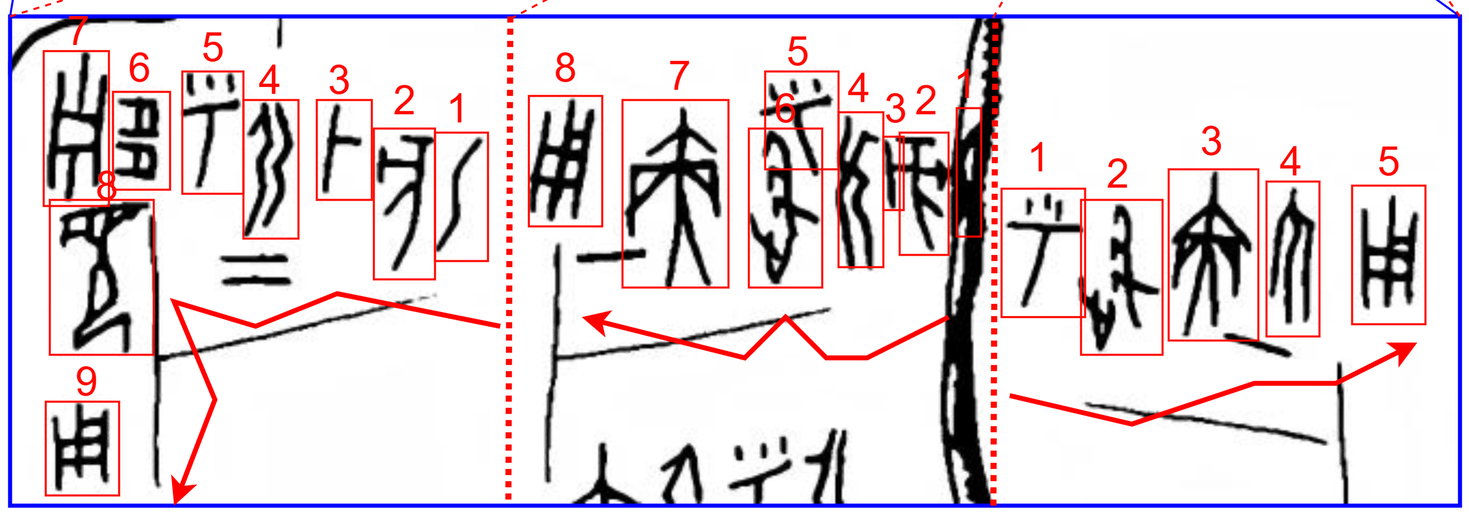
花〇〇二五五　癸子歲癸子羽一
花〇〇二五五　癸子歲癸子羽一
花〇〇二五五　壬戌…口…
花〇〇二五四　…庚…
花〇〇二五五　甲寅卜弘宜丁
花〇〇二五五　甲寅卜弘言來自西祖乙廿又伐
花〇〇二五五　弘乎弓ㄓ
花〇〇二五五　乎徵ㄓ不用
花〇〇二五五　乙亥卜弘乎弓ㄓ而
花〇〇二五五　乙亥卜弘乎多　而見用
花〇〇二五五　丁丑歲匕庚一牝子往滿御〔興〕
花〇〇二五五　戊寅卜舟曰告昌弗虞永
花〇〇二五五　壬卜…
花〇〇二五六　壬卜三日雨至
花〇〇二五六　壬卜五日雨至
花〇〇二五六　己…牛羊…庚…
花〇〇二五六　…小宰…匕庚
花〇〇二五六　…丁…其…
花〇〇二五六　…于庚…
花〇〇二五六　…乙卜…
花〇〇二五六　…己卜…

五六三九

Characters



Sectences



Rubbing Characters

Facimile Characters

Sectence Reading Order

(a)

王

吉

日

Rubbing Character Dataset

(b)

Facsimile Character Dataset

(c)

Objection Detected Dataset

(d)

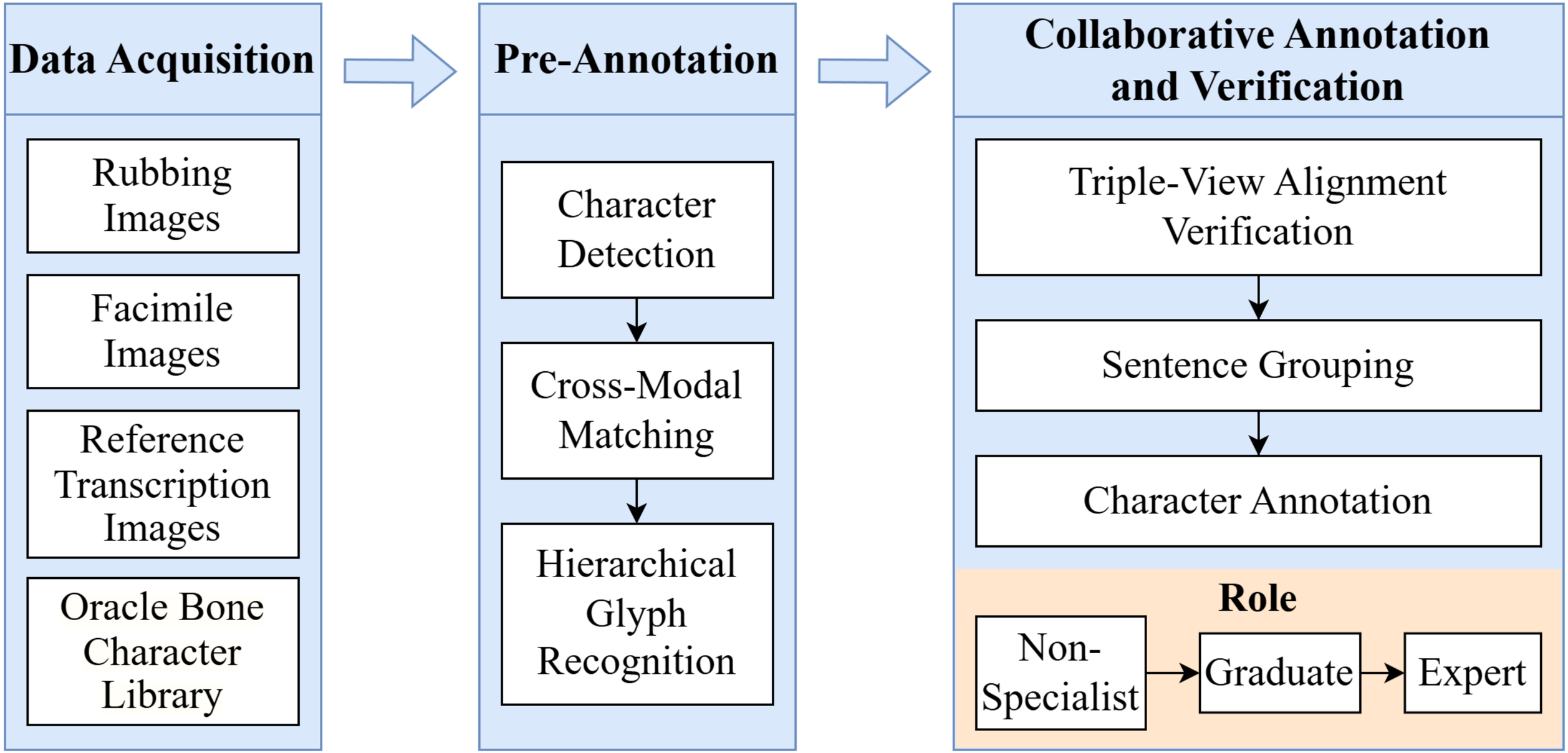
庚 卜 史 贞 王 東 吉 燕

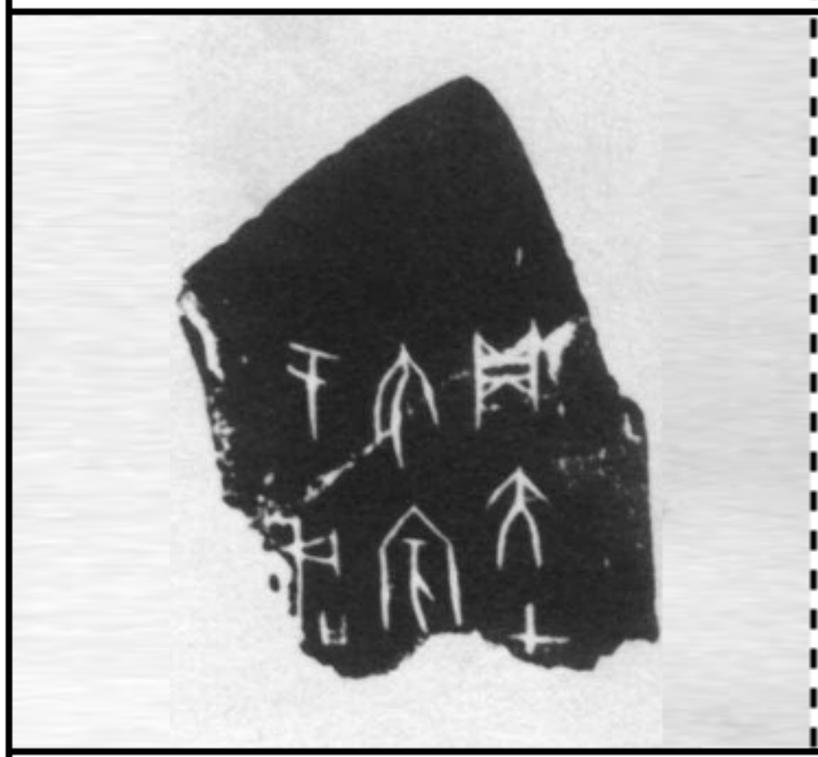
之 日

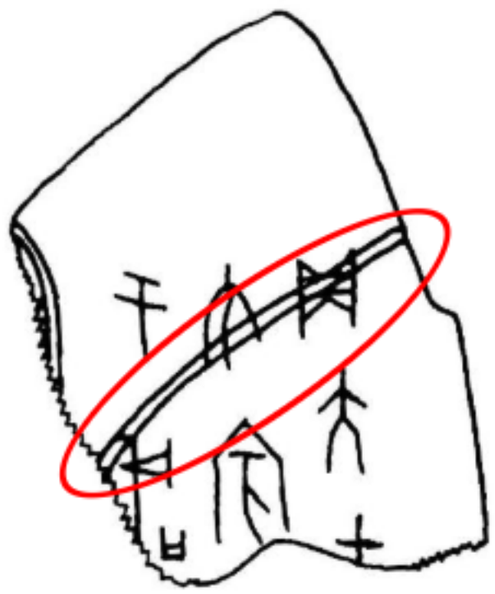
一

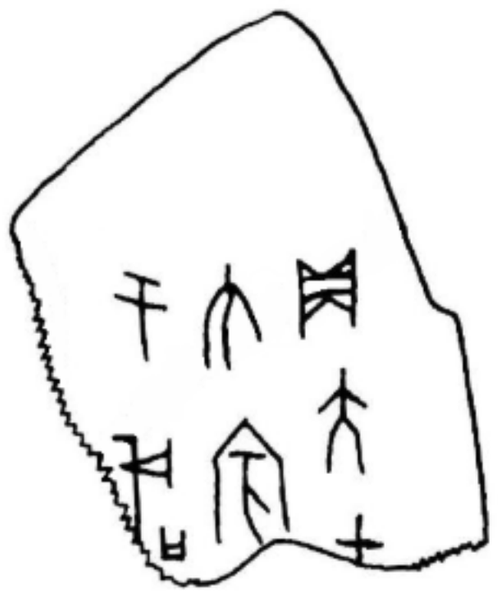
吉 燕 八月

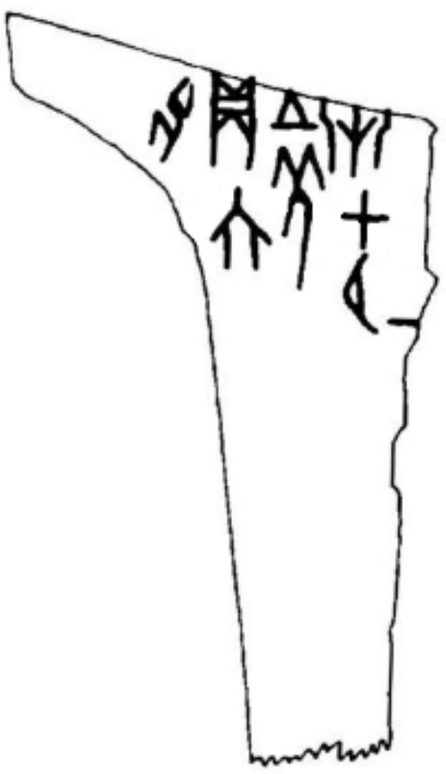
OBIMD


Data Acquisition
Rubbing Images
Facimile Images
Reference Transcription Images
Oracle Bone Character Library
Pre-Annotation
Character Detection
Cross-Modal Matching
Hierarchical Glyph Recognition
Collaborative Annotation and Verification
Triple-View Alignment Verification
Sentence Grouping
Character Annotation
Role
Non-Specialist
Graduate
Expert





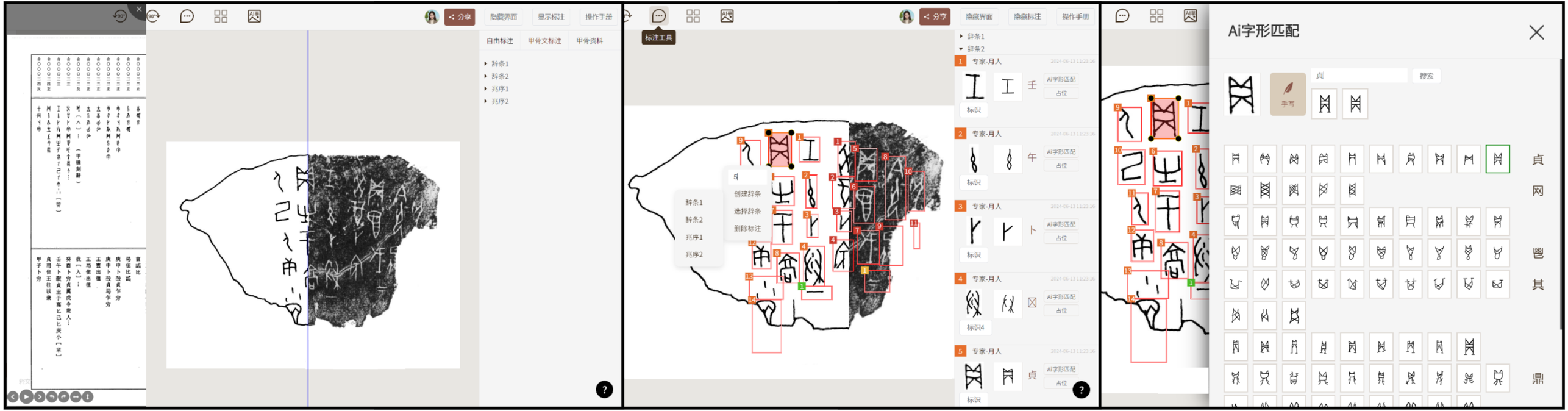




(a) Rubbing

(b) Original Facsimile

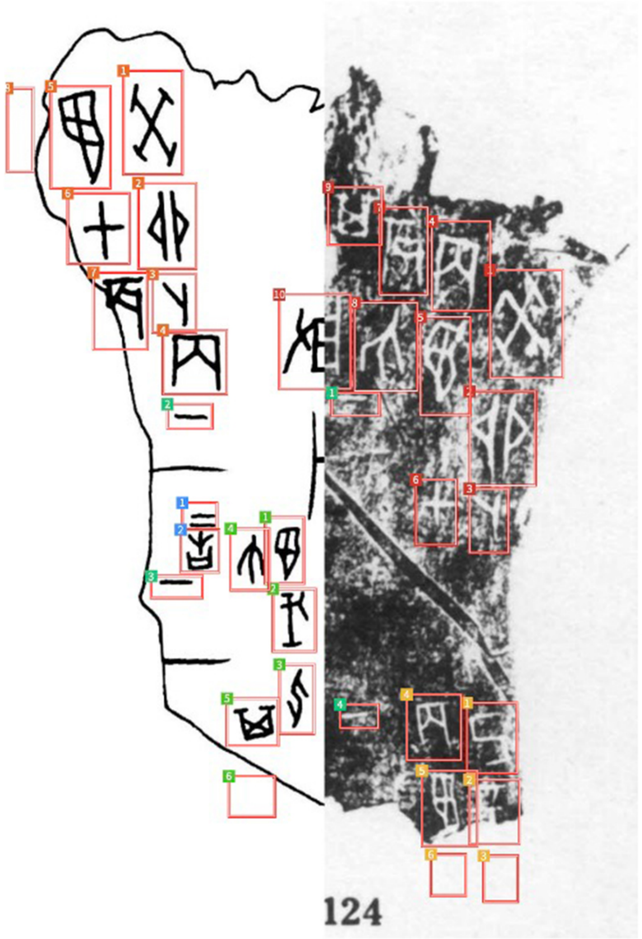
(c) Redrawn Facsimile



(a) Triple-View Alignment Verification (b) Sentence Grouping (c) Character Annotation

Facsimile

Rubbing



Bounding Boxes、
Inscription Groups

Inscription Group 1

	OBI	CC	TRA
1			癸
2			卯
3			卜
4			丙
5			翌
6			甲
7			辰
8			不
9			其
10			啟

Reading Sequences

Inscription Group 2

	OBI	CC	TRA
1			癸
2			卯
3			卜
4			丙
5			翌
6			甲
7			辰
8		...	

Reading Sequences

Inscription Group 3

	OBI	CC	TRA
1			己
2			酉
3		...	
4			丙
5			翌
6		...	

Reading Sequences

OBI: Oracle Bone Inscriptions
CC: Character Categories
TRA: Transcriptions

Inscription Group 4

	OBI	CC	TRA
1			翌
2			戊
3			申
4			不
5			其
6		...	

Reading Sequences

Oracle Outcome

OBI	CC	TRA
	二	二
	刃	告

Oracle Sequence

OBI	CC	TRA
	一	一
	一	一
	一	一
	一	一