Explainable Oracle Bone Script Recognition via Multimodal Pictographic Reasoning

Yin Wu[1], **Zhengxuan Zhang**[1], **Jiayu Chen**[1], **Chang Xu**[1], **Yuyu Luo**[1, 2], **Nan Tang**[1, 2], **Hui Xiong**[1, 2*]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
ywu450@connect.hkust-gz.edu.cn, zzhang393@connect.hkust-gz.edu.cn, jchen161@connect.hkust-gz.edu.cn,
cxu475@connect.hkust-gz.edu.cn, yuyuluo@hkust-gz.edu.cn, nantang@hkust-gz.edu.cn, xionghui@hkust-gz.edu.cn

Abstract

Oracle Bone Script, East Asia's earliest mature writing system from over 3,500 years ago, encodes ancient cognition through visual metaphors, yet remains largely undeciphered and inaccessible, severing modern society from its cultural roots. Traditional AI methods, while accurate in classification, treat glyphs as opaque data, neglecting their pictographic essence and failing to foster public understanding—exacerbating a heritage crisis amid linguistic evolution. We pioneer a paradigm shift toward AI-driven cultural democratization, introducing **OracleVis**, the first human-validated multimodal dataset of glyph-image-explanation triplets, curated through expert collaborations to overcome data scarcity, bias, and incompleteness in archaeological sources. Building on this, **OBS-VM**, an explainability-centric multimodal large language model fine-tuned on Qwen2-VL-7B, models pictographic reasoning by balancing semantic fidelity with interpretive transparency, transforming black-box predictions into cognition-aligned narratives. Rigorous evaluations, including benchmarks and a user study with 24 non-experts, reveal our system's superiority: it outperforms GPT-4o in **pictographic rationality (3.79 vs. 3.58 in human evaluation)** and achieves a **35.3% relative improvement in recognition accuracy**, while interactive learning boosts knowledge gains (+5.5 vs. +1.7), interest (+1.9 vs. +0.4), and confidence (+2.0 vs. +0.3) over static methods. This work illuminates AI's potential to bridge ancient wisdom and contemporary audiences, redefining heritage preservation as an inclusive, socially impactful endeavor that turns cultural alienation into enlightened engagement.

Introduction

Oracle bone script (OBS), dating from 1600-1046 BCE, represents East Asia's earliest mature writing system and a foundational pillar of Chinese civilization. This pictographic system offers a unique window into ancient cognition (Chiang 2006; Da WEIH 2023), where abstract concepts were systematically encoded through visual metaphors. As shown in Figure 1, Physical objects (for "ax"), dynamic actions (for "horse riding"), and intangible ideas (for "fire") were transformed into structured glyphs that preserve intuitive human perception patterns. These visual metaphors function as cognitive bridges, revealing the tools, social structures, and conceptual frameworks that defined ancient society.

Figure 1: OBS as a visual archive of ancient civilization. These pictographs are more than just characters; they are direct records of the material culture, social activities, and natural concepts of the Shang Dynasty. Each glyph serves as a window into the cognition and culture of ancient people.

The Social Disconnect: A Cultural Heritage Crisis. Despite its monumental historical significance, OBS faces a severe social crisis that extends far beyond academic boundaries. After millennia of linguistic evolution, approximately 60% of discovered characters remain undeciphered (Wang and Deng 2024), while even the successfully interpreted glyphs demand specialized paleographic expertise for meaningful comprehension. This creates a profound cognitive disconnect between contemporary society and its cultural origins – a barrier that not only impedes scholarly research but, more critically, severs the connection between ordinary citizens and their ancestral heritage. The resulting cultural alienation undermines educational objectives, restricts public participation in heritage preservation, and perpetuates an elite-dominated understanding of foundational cultural knowledge.

The Core Challenges: Why is This Heritage Silent? Current computational approaches to OBS, while achieving technical success in classification tasks (Zhen, Wu, and Liu 2024; Wang et al. 2024d; Yang et al. 2024), fundamentally misalign with the social impact goals essential for cultural heritage democratization. The challenges span both data and methodology, creating structural barriers to meaningful public engagement:

*Corresponding author

(C1) **Data modeling crisis: Missing pictographic foundations.** Existing OBS datasets fundamentally lack the pictographic elements essential to the script's nature. Current collections contain only isolated glyph images without corresponding real-world pictographic references or expert interpretations of the visual metaphors underlying character formation (Guan et al. 2024; Da WEIH 2023). This absence strips away the core logic that connects ancient symbols to tangible objects and concepts, reducing rich cultural artifacts to sterile pixel data that cannot convey the intuitive visual reasoning that made these characters accessible to ancient peoples.

(C2) **Methodological barrier: Black-box reasoning limitations.** Even with high classification accuracy (ResNet-50 achieves 94.3% on known OBS (Wang et al. 2024c)), current methodological approaches provide minimal educational value. Black-box models cannot articulate the pictographic reasoning that connects glyphs to their real-world origins—the very insight needed for public understanding. The methodological challenge shifts from optimizing "classification accuracy" to developing "interpretable reasoning frameworks" that can explain, in human-readable terms, how visual metaphors encode meaning. Without this explanatory capability, AI remains confined to expert circles rather than serving as a democratizing force for cultural heritage.

Our Contribution: AI as a Bridging Tool for Cultural Democratization. The rapid advancement of AI technologies, particularly Multimodal Large Language Models (MLLMs) that facilitate seamless interactions through text and images, has unveiled unprecedented opportunities for cultural democratization. Inspired by this potential, we posit that these challenges can be effectively confronted from two interconnected fronts: innovative data construction and advanced AI methodologies. As conceptually illustrated in Figure 2, our work aims to transform the user experience from one of confusion to one of enlightenment by re-centering the analytical process on pictographic reasoning.

To this end, we introduce **OracleVis**, the first human-validated multimodal dataset for pictographic OBS analysis. Drawing from verified archaeological records and expert knowledge bases, it features "glyph-image-textual explanation" triplets with an average explanation length of 1,096.34 characters. This dataset provides a systematic resource linking OBS symbols to real-world semantics, enabling public access to character meanings and underlying reasoning. Building on **OracleVis** to unlock its full explanatory power for non-experts, we developed **OBS-VM**, a domain-adapted MLLM centered on explainability. Using a two-stage fine-tuning—first on OBS textual descriptions for semantic adaptation, then on real-world images for pictographic capture—**OBS-VM** generates cognition-aligned explanations. Users can input OBS characters for comprehensive, understandable interpretations, surpassing GPT-4o in **pictographic rationality (3.79 vs. 3.58)** and achieving a **35.3% relative improvement in recognition accuracy**.

Together, **OracleVis** and **OBS-VM** form a novel AI-driven pathway for popularizing ancient scripts, integrating social needs like explainability and accessibility into data, model, and evaluation frameworks. This approach addresses humanities challenges and demonstrates AI's transformative social impact through cultural heritage democratization.

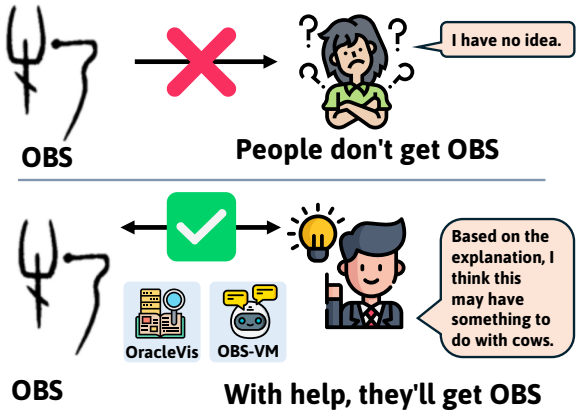

Figure 2: Conceptual illustration of the paradigm shift enabled by our approach. (Top) Without pictographic context, OBS is an incomprehensible 'black box' for non-experts, creating cultural and cognitive barriers. (Bottom) Our framework bridges this gap by integrating a pictographic dataset (**OracleVis**) and explainable model (**OBS-VM**), revealing the visual logic behind the characters and transforming confusion into understanding.

Our contributions can be summarized as follows:

- **Cultural heritage preservation:** We introduce **OracleVis**, the first human-validated multimodal dataset for analyzing OBS pictographic logic, providing valuable digital infrastructure for cultural transmission in the digital age and establishing systematic foundations for ancient Chinese script research.
- **Education and public engagement:** We propose **OBS-VM**, an explainability-centered multimodal large language model that not only surpasses GPT-4o in **pictographic rationality (3.79 vs. 3.58)** and achieves a **35.3% relative improvement in recognition accuracy**, but more importantly, demonstrates capability in helping non-specialists understand OBS character formation logic, opening new directions for AI applications in historical education and cultural popularization.
- **New paradigm for social impact:** Overall, this project exemplifies how AI can effectively address complex problems in humanities and social sciences by deeply integrating social needs into data construction, model design, and evaluation processes, demonstrating artificial intelligence's significant potential for social impact through cultural heritage democratization.

Related Works

Recent advancements in machine learning methodologies have been increasingly applied to the study of OBS. Research in this domain predominantly focuses on data pre-

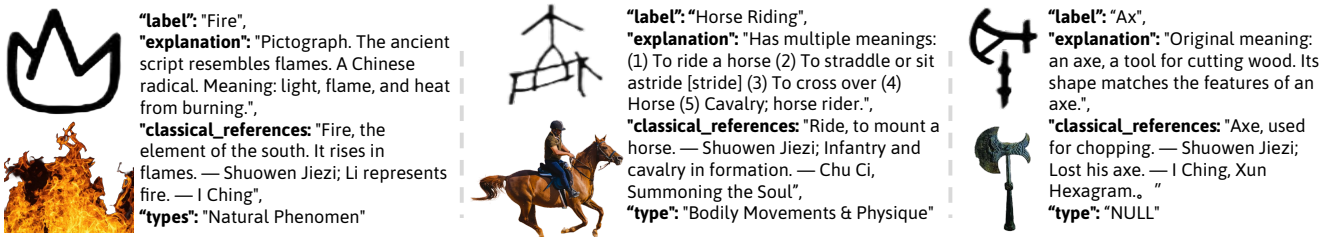

Figure 3: Three example data in **OracleVis**. Due to space constraints, we show here a portion of the text data in each example. For better readability, we have translated the data into English. For the original Chinese data, please refer to the appendix.

processing, information retrieval, pattern recognition, and inscription decipherment. Chen et al. (2024) introduced an innovative framework called OBI-bench. This benchmark assesses the efficacy of computational models in addressing key challenges: the identification, classification, retrieval, and decipherment of Oracle Bone inscriptions. These areas encompass the primary applications of machine learning in OBS. Existing datasets related to OBS encompass extensive resources, including high-fidelity rubbings, digitized handwriting samples, and corresponding character notations. Advanced graph detection techniques, exemplified by Zhen, Wu, and Liu (2024), effectively extract Oracle Bone characters from physical fragments. Concurrently, graph denoising algorithms, as delineated in Wang et al. (2024d), and the two-stage (coarse-to-fine) network proposed by Yang et al. (2024) enhance data quality through efficient OBS image inpainting. Such methodologies have significantly elevated the quality of datasets, exemplified by the HUST-OBI dataset introduced in Wang et al. (2024c), an open-access dataset that aggregates images from scholarly texts, books, online repositories, and other databases, including HWOBC (Li et al. 2020). Building upon HUST-OBI, we further integrate comprehensive annotations, including semantic explanations, character classifications, additional semantic contexts, and corresponding real-world imagery.

Despite these data-processing advancements, oracle character recognition continues to grapple with three critical challenges: inherent variability in writing styles, the paucity of annotated data, and the degraded quality of scanned imagery. Addressing these impediments, scholars (Li et al. 2024) have employed various sophisticated methodologies. Convolutional Neural Network (CNN)-based approaches have been utilized to elucidate the relationships between simplified Chinese characters and Oracle Bone inscriptions (Fu et al. 2022; Mai, Penava, and Buettner 2024). Additionally, contrastive learning frameworks, combining unsupervised ResNet-50 with supervised REPVGG, have demonstrated efficacy in retrieving analogous Oracle Bone images (Weng et al. 2024). Beyond direct character recognition, innovative approaches have advanced the field by deconstructing Oracle Bone characters into their radical components (Hu et al. 2024; Wang et al. 2024b). Recent work on integrating typographic and stylistic variations across historical periods (Wu et al. 2024) and the application of diffusion models (Guan et al. 2024) has further enhanced the recognition and decipherment of OBS. However, these approaches require large amounts of labeled data, which is time- and resource-consuming.

Multimodal Large Language Models

The evolution of language models has undergone a remarkable transformation, particularly through the development of Large Language Models (LLMs) that employ advanced neural network architectures like Transformers (Vaswani 2017) and leverage extensive pre-training datasets. State-of-the-art models such as GPT (Radford 2018) demonstrate exceptional capabilities in capturing nuanced linguistic patterns and generating contextually aware responses, driving advancements across diverse application domains (Liu et al. 2025; Sheng et al. 2024). This progress has been further accelerated by Parameter-Efficient Fine-Tuning (PEFT) techniques (Hu et al. 2021), which optimize training efficiency while maintaining model performance.

OracleVis Construction

In this study, we construct a high-quality multimodal OBS dataset, named **OracleVis**, which integrates images, textual descriptions, and background knowledge, as shown in Figure 3. This dataset is designed to train MLLMs for OBS recognition and interpretation. Each annotated sample in **OracleVis** is structured as:

$$\mathcal{D} = \{(i_{o,k}, i_{r,k}, \mathbf{t}_k)\}_{k=1}^{K}$$

where K is the total number of instances, i_o represents the OBS image, i_r is a real-world object image, and $\mathbf{t} = (t_d, t_p, t_h, t_c)$ denotes expert-annotated textual descriptions. Specifically, t_d is the modern Chinese character corresponding to the OBS glyph (the "label" in Figure 3), t_p describes how the glyph's structure visually represents a real-world object (the "explanation"), t_h provides historical occurrences of the glyph in ancient texts (the "classical_references"), and t_c indicates the semantic classification as defined in Multi-function Chinese Character Database[6] (the "type").

Data Sources. The **OracleVis** dataset is constructed from multiple authoritative sources to ensure diversity and reliability [1]. As summarized in Table 1, it consists of three primary components. The OBS images are collected from established databases, scanned ancient texts, and expert-created standardized illustrations. Additionally, real-world object images are

[1]The collected Original data: https://drive.google.com/drive/folders/1bwEQnHB7kD37C-S5ALkBfq5Cks-yu34_?usp=sharing

Data Category	Books Source	Websites Source	Databases Source
OBS Images	Liu, Hong, and Zhang (2009); Liu and Liu (2019)	GuoXueDaShi[2]; YinQiWenYuan[3]	Wang et al. (2024c); Li et al. (2020)
Real-World Object Images	-	Google Image[4]; Pinterest[5]	-
Human-Annotated Text	Liu, Hong, and Zhang (2009); Liu and Liu (2019)	Archaic Sound Micro Mirror[6]; Multi-function Chinese Character Database[7]	Wang et al. (2024c)

Table 1: Sources of different data categories.

manually selected to align with the pictographic characteristics of OBS. These images were retrieved through a structured search process to maintain accuracy and relevance. The expert-annotated textual descriptions are derived from reputable linguistic and archaeological resources, including the HUST-OBS dataset (Wang et al. 2024c), the Multi-function Chinese Character Database, and the Archaic Sound Micro. Each annotation is verified by experts to incorporate insights from historical linguistics and archaeological research.

Two-stage Data Refinement. To enhance the **pictographic emphasis** in textual explanations, we implement a two-stage refinement framework combining automated cross-modal alignment with expert-guided validation, as shown in Figure4. This process addresses the limitations of initial annotations that lack granular visual-semantic analysis crucial for OBS interpretation.

The first stage employs cross-modal semantic anchoring. We utilize GPT-4o, a state-of-the-art vision-language model, to generate initial pictographic explanations by establishing glyph-referent mappings. For each OBS glyph-image pair $(i_o, i_r) \in \mathcal{I}_{\text{OBS}} \times \mathcal{I}_{\text{Real}}$, the model generates:

$$t_p^{(0)} = \mathcal{M}(i_o, i_r),$$

where $\mathcal{M}$ denotes the teacher model's cross-modal mapping function. This automated phase anchors visual features (e.g., stroke patterns in i_o) to semantic attributes of real-world objects in i_r.

The second stage is expert-guided iterative refinement. To mitigate potential **semantic misalignments** in model-generated outputs, we implement a human-in-the-loop refinement Mechanism. Experts systematically review and refine the AI-generated explanations $t_p^{(0)}$, ensuring that the pictographic relationships are clearly articulated and historically contextualized. Domain experts iteratively refine $t_p^{(0)}$ through:

$$t_p^* = \mathcal{V}_{\text{expert}}(t_p^{(0)}, \mathcal{C}_{\text{ref}}),$$

where $\mathcal{C}_{\text{ref}} = \{t_d, t_h, t_c\}$ represents reference annotations from the original dataset. Experts validate three key aspects: (1) **Structural congruence** between glyph components and referent object; (2) **Historical consistency** with archaeological evidence; (3) **Linguistic precision** in semantic categorization.

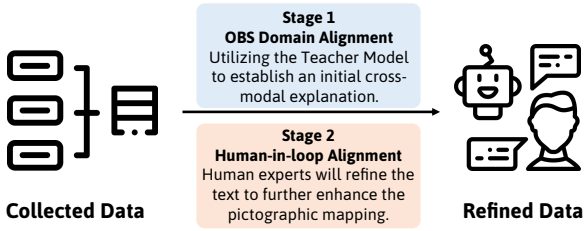

Figure 4: Two-stage refinement process for enhancing pictographic interpretation in **OracleVis**. Stage 1 (blue): GPT-4o generates an initial cross-modal explanation by aligning an OBS glyph with a corresponding real-world referent. Stage 2 (orange): Human experts iteratively refine the AI-generated text to improve the accuracy of pictographic mapping.

In this two-stage process, the verified explanations serve as the final ground-truth annotations used during both model training and evaluation. When the model-generated output is sufficiently accurate, coherent, and semantically aligned, it is accepted without modification. However, when semantic misalignments, structural inconsistencies, or lack of historical grounding are identified, experts revise the explanation to ensure an alignment between the glyph's features, its real-world referent, and its historical semantics.

Two-Stage Model Fine-tuning

To effectively adapt a pre-trained MLLM for the specialized task of OBS interpretation, we leverage **Qwen 2-VL-7B** (Wang et al. 2024a) as our base model. This choice is motivated by its demonstrated proficiency in vision-language tasks and its extensive pre-training corpus, which includes substantial Chinese data crucial for understanding the OBS context. Given the significant divergence between the linguistic structures of ancient Chinese texts and modern language—including differences in grammar, vocabulary, and writing conventions—direct training of a pre-trained MLLM

[2]https://www.guoxuedashi.net/
[3]https://jgw.aynu.edu.cn/
[4]https://images.google.com
[5]https://nl.pinterest.com/
[6]http://www.kaom.net/index.php
[7]https://humanum.arts.cuhk.edu.hk/Lexis/lexi-mf/

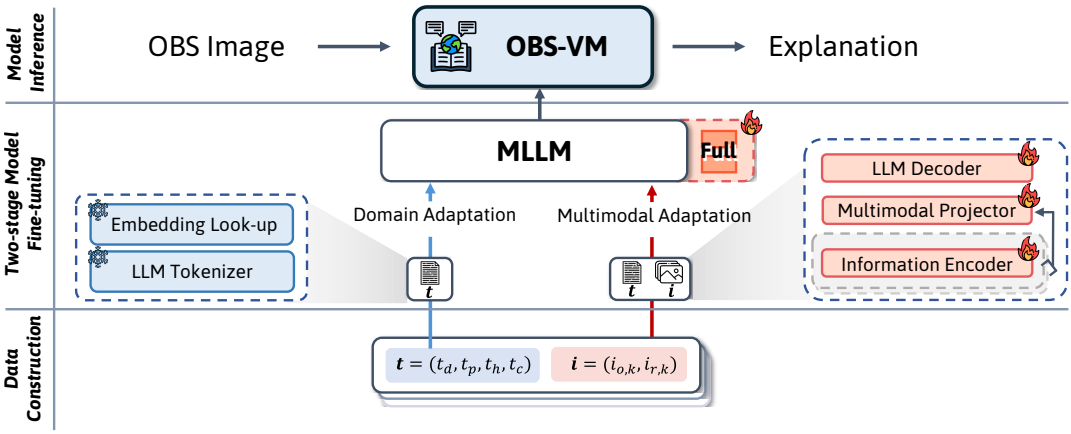

Figure 5: Overview of our two-stage fine-tuning strategy for **OBS-VM**. Stage 1 (Domain Adaptation) adapts the language model component to OBS textual annotations while freezing visual modules. Stage 2 (Multimodal Adaptation) integrates visual inputs (i_o and i_r) and performs full fine-tuning to establish robust visual-semantic alignment for generating comprehensive explanations t derived from $\mathbf{t} = (t_d, t_p, t_h, t_c)$.

on OBS data often results in suboptimal performance. To address this challenge, we propose a two-stage fine-tuning strategy that first enables the model to internalize the unique linguistic system of ancient texts before integrating visual information. The primary aim of the first stage is to help the model acquire a deeper understanding of the stylistic and semantic patterns prevalent in historical annotations, thereby laying a solid foundation for robust multimodal alignment in the subsequent stage.

Stage 1: Domain Adaptation. The primary objective of this initial stage is to tune the language model component of the MLLM to the specific linguistic style, terminology, and semantic structure inherent in OBS textual annotations. We utilize the textual components $\mathbf{t} = (t_d, t_p, t_h, t_c)$ from our dataset $\mathcal{D}$. Crucially, during this phase, we **freeze the parameters of the visual encoder and the cross-modal attention layers**. This focused approach prevents the model from prematurely altering its learned visual representations while concentrating computational resources on adapting the language processing modules (specifically, the LLM's self-attention and feed-forward layers). The model is trained using a standard auto-regressive language modeling objective, minimizing the negative log-likelihood of generating the target text sequence $\mathbf{t}$ given the (unused) visual inputs. This ensures the model learns to generate coherent and contextually relevant textual descriptions based on the patterns observed in the annotations. Let Φ_0 be the initial pre-trained parameters. The objective is:

$$\min_{\Theta_{LM}} \sum_{(\cdot,\cdot,\mathbf{t})\sim\mathcal{D}} -\log P(\mathbf{t}; \Phi_{0,visual}, \Theta_{LM}),$$

where Θ_{LM} represents the trainable language model parameters, and $\Phi_{0,visual}$ represents the frozen visual and cross-modal parameters. Let Φ_1 denote the parameters after this stage.

Stage 2: Multimodal Adaptation. Building upon the domain-adapted language model from Stage 1, the second stage integrates the visual modality to establish profound visual-semantic connections. This stage leverages the complete data structure from $\mathcal{D}$, incorporating the OBS glyph image i_o, the real-world reference image i_r, and the target textual explanation t (representing the comprehensive output derived from $\mathbf{t}$). Unlike the previous stage, we perform **full fine-tuning**, unfreezing all model parameters ($\Theta = \{\Theta_{visual}, \Theta_{cross-modal}, \Theta_{LM}\}$). This allows for end-to-end optimization, enabling the model to learn complex cross-modal alignments between the visual features of the glyph (i_o), the corresponding real-world object (i_r), and the associated semantic information (t). The objective is to maximize the conditional likelihood of generating the accurate and comprehensive textual explanation t given both visual inputs:

$$\max_{\Theta} \sum_{(i_o,i_r,t)\sim\mathcal{D}} \log P(t \mid i_o, i_r; \Phi_1, \Theta),$$

where Φ_1 represents the parameters initialized from Stage 1, and Θ denotes all trainable parameters of the model. The use of **expert-verified descriptions** within $\mathcal{D}$ is critical here, providing a high-quality supervision signal that guides the model towards generating explanations that are not only semantically coherent but also maintain historical and linguistic accuracy, effectively bridging the gap between the pictographic form and its meaning.

Experiments

In this section, we conducted experiments to evaluate the effectiveness of **OracleVis** and **OBS-VM**. Specifically, we aim to answer the following evaluation questions:

Q1: How effective is **OracleVis** in improving a model's pictographic OBS interpretation and recognition capabilities?

Q2: To what extent can **OBS-VM**'s explanations capture the pictographic logic and deep cultural context of OBS,

Models	Human Evaluation			GPT Evaluation			Accuracy	
	PR	**SR**	**RD**	**PR**	**SR**	**RD**	**Top-1**	**Top-5**
OBSD	/	/	/	/	/	/	0.13	/
GPT-4o	3.58	4.29	4.03	4.37	4.29	**4.50**	0.17	0.24
Claude-3.5-Sonnet	3.63	4.23	3.98	4.29	4.07	4.43	0.14	0.18
GLM-4V	2.87	3.45	3.77	4.13	3.87	3.73	0.07	0.11
Qwen2-VL-7B	2.37	2.86	2.65	3.65	3.29	2.89	0.10	0.14
OBS-VM$_{\text{w/o}}$ (MA)	2.31	2.53	3.52	3.99	3.67	3.56	0.13	0.16
OBS-VM	**3.79**	**4.33**	**4.19**	**4.40**	**4.41**	4.33	**0.23**	**0.33**

Table 2: Performance comparison of **OBS-VM** and baseline models on OBS explanation and recognition tasks. The table presents human and GPT-based evaluations across three explanation quality metrics—PR, SR, and RD—as well as recognition accuracy (Top-1 and Top-5). **OBS-VM** achieves the highest scores across **key** metrics, demonstrating the effectiveness of domain-specific fine-tuning. Ablation results (**OBS-VM**$_{\text{w/o}}$ (MA)) highlight the importance of multimodal adaptation in improving both explanation quality and recognition accuracy.

when compared to those generated by humans?
Q3: Whether interactive learning with **OBS-VM** could significantly enhance non-specialists' understanding, interest, and confidence compared to traditional methods.

Experimental Setup

Implementation Details. We use Qwen2-VL-7B (Wang et al. 2024a) as the base MLLM, leveraging its strong Chinese language capabilities from training on extensive Chinese corpora, ideal for ancient scripts like OBS. Training follows a two-stage fine-tuning: domain adaptation freezes the visual encoder and multimodal projection layers, updating only text embedding and attention layers with OBS glyphs and descriptions. Multimodal adaptation unfreezes the full model, incorporating OBS glyphs and real-world images for pictographic understanding. Batch sizes are 64 (domain) and 128 (multimodal). We use AdamW optimizer with 0.1 weight decay, linear warmup from 1e-8 to 5e-4 learning rate, and cosine decay. Training uses 8 NVIDIA H100 (80GB) GPUs.

Dataset. Training uses 2,396 samples from **OracleVis**, covering 289 distinct OBS glyphs. Domain adaptation employs the original dataset; multimodal adaptation uses a ShareGPT-formatted version for better text alignment and consistency.

Evaluation uses a 1,000-sample test set with pictographic examples, including seen glyphs (same 289 types but different styles) for generalization testing and unseen glyphs for zero-shot recognition assessment.

Baselines. We compare **OBS-VM** with leading multimodal models and OBS-specific ones: GPT-4o (Achiam et al. 2023) and Claude 3.5 Sonnet (via API) for broad multimodal strength; GLM-4V (GLM et al. 2024) for Chinese multimodal tasks (via API); base Qwen2-VL-7B to measure fine-tuning gains; and OBSD (Guan et al. 2024), a diffusion-based OBS recognition model (provided with our training OBS images and aligned modern Chinese images). OBSD lacks explanations, so it is excluded from explainability metrics.

Performance Evaluation. Due to high expert annotation costs, we use hybrid human-GPT evaluation. Eight institutional volunteers (undergraduate-level, no paleography expertise) simulate non-specialist users. Each sample is scored by three annotators; we average scores with Krippendorff's alpha of 0.82 for reliability.

Explanations are rated on a 5-point Likert scale (5 = highest, 1 = lowest) for:

- **Pictographic Rationality (PR)**: Clarity in describing OBS glyph's pictographic features.
- **Semantic Rationality (SR)**: Semantic accuracy aligning with modern Chinese meanings.
- **Readability (RD)**: Coherence and ease of understanding.

For recognition, models output top-5 predictions per glyph, measured by:

- **Top-1 Accuracy**: Percentage where top prediction is correct.
- **Top-5 Accuracy**: Percentage where correct answer is in top 5.

OBSD uses PaddleOCR for generated glyph recognition, following Guan et al. (2024).

Performance Comparison

To evaluate the performance of **OBS-VM** in generating pictographic explanations for OBS, we benchmarked it against a suite of SOTA multimodal models, including GPT-4o, and Qwen2-VL-7B. The evaluation centered on explanation quality metrics (PR, SR, RD) and recognition accuracy, with results detailed in Table 2.

OBS-VM demonstrates a superior ability to interpret the pictographic logic of OBS. In the human evaluation, it achieved the highest Pictographic Rationality (PR) score of 3.79. This result not only surpasses the leading SOTA model, GPT-4o (3.58), but also dramatically outperforms its base model, Qwen2-VL-7B (2.37). This exposes a potential weakness in general-purpose models: their tendency to "hallucinate" explanations by applying knowledge of modern characters to ancient scripts, thus failing to capture authentic visual-semantic mappings.

Furthermore, **OBS-VM** attained the highest recognition accuracy, achieving a Top-1 score of 23.0% (0.23). This represents a significant 35.3% improvement over GPT-4o (17.0%

Metric	Group	Pre-test (Mean ± SD)	Post-test (Mean ± SD)	Gain / Final Score
Objective Metric				
Understanding Score (/10)	Experimental (N=12)	1.6 ± 0.8	7.1 ± 1.1	**+5.5**
	Control (N=12)	1.7 ± 0.7	3.4 ± 0.9	+1.7
Subjective Metrics				
Interest (/5)	Experimental (N=12)	2.7 ± 0.6	4.6 ± 0.5	**+1.9**
	Control (N=12)	2.8 ± 0.7	3.2 ± 0.6	+0.4
Confidence (/5)	Experimental (N=12)	2.4 ± 0.5	4.4 ± 0.4	**+2.0**
	Control (N=12)	2.5 ± 0.6	2.8 ± 0.7	+0.3
Material Helpfulness (/5)	Experimental (N=12)	-	4.8 ± 0.3	**4.8**
	Control (N=12)	-	2.5 ± 0.8	2.5

Table 3: User Study Results: Model-Assisted vs. Traditional Learning. The experimental group refer the information powered by **OBS-VM**, while the control group accessed static web resources (GuoXueDaShi and YinQiWenYuan) and could perform their own searches. Understanding Scores are out of 10, while subjective scores are on a 5-point Likert scale. Gain is calculated as (Post-test - Pre-test). N=12 participants were randomly assigned to each group.

Methods	PR	SR	RD
Human	4.20	4.56	4.30
OBS-VM	3.78	4.20	4.28

Table 4: Performance comparison between human interpretation and **OBS-VM**.

or 0.17). Similarly, **OBS-VM** also led in Semantic Rationality (SR) with a human evaluation score of 4.33 (vs. GPT-4o's 4.29). This superior performance across rationality and accuracy suggests that our model's domain-specific fine-tuning successfully prioritizes the core pictographic reasoning essential for correct interpretation.

Comparison with Human Explanation

To benchmark **OBS-VM** against human interpretive capabilities, we compared its outputs against explanations from five native Chinese speakers for a shared set of 50 glyphs. While **OBS-VM** produced explanations with near-human linguistic fluency, they lacked the semantic depth and pictographic rationality of the human-generated accounts (Table 4).

Specifically, human interpretations scored significantly higher on PR and SR. This result indicates that human reasoning integrates a broader reservoir of cultural background, historical context, and implicit knowledge—elements not explicitly encoded in our dataset. In stark contrast, the negligible difference in RD reveals that **OBS-VM** excelled at replicating the linguistic style and structure of the training explanations. The model thus reproduces the form of an authentic explanation but fails to capture its full interpretive substance.

User Study: Evaluating OBS-VM as a Learning Aid

To empirically validate our claim that **OBS-VM** can serve as an effective bridge for non-specialists to understand the pictographic logic of OBS, we conducted a user-centric study. We employed a pre-test/post-test control group design with 24 non-expert participants. The experimental group used an interactive interface powered by **OBS-VM** to explore OBS glyphs, while the control group used static documents with traditional textual explanations.

As shown in Table 3, the results provide strong support for our hypotheses. In our objective evaluation, the model-assisted group demonstrated a substantially higher **knowledge gain score** (+5.5) compared to the control group (+1.7). This indicates that interacting with the model's visual and textual explanations equips users with a superior ability to decipher the logic of unseen OBS characters.

Subjective feedback further reinforces this finding. The experimental group reported a dramatic increase in both **interest** (+1.9 vs. +0.4) and **confidence** (+2.0 vs. +0.3) after the intervention. Furthermore, they rated the helpfulness of the interactive tool (4.8/5) nearly twice as high as the control group rated the static materials (2.5/5). This suggests that the model does not just convey information, but also fosters a more engaging and empowering learning experience. This user-centric evidence validates that **OBS-VM** is an effective tool for promoting the democratization and public engagement of cultural heritage.

Conclusion

This work bridges the cultural gap between modern society and ancient text, East Asia's earliest writing system, via an AI-driven approach transforming opaque glyphs into accessible heritage. We present **OracleVis**, a human-validated multimodal dataset of glyph-image-explanation triplets encoding pictographic logic, and **OBS-VM**, a MLLM for cognition-aligned interpretations. Shifting from black-box process to interpretable reasoning, it overcomes barriers limiting access to experts. User studies show interaction boosts non-specialists' knowledge, interest, and confidence over traditional methods. By integrating social needs into data and model design, our framework democratizes cultural heritage, providing a replicable approach for revitalizing other ancient legacies.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (Grant No.2023YFF0725001), in part by the National Natural Science Foundation of China (Grant No.92370204), in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No.2023B1515120057), in part by the Key-Area Special Project of Guangdong Provincial Ordinary Universities (2024ZDZX1007), in part by the Education Bureau of Guangzhou.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen, Z.; Chen, T.; Zhang, W.; and Zhai, G. 2024. OBI-Bench: Can LMMs Aid in Study of Ancient Script on Oracle Bones? *arXiv preprint arXiv:2412.01175*.

Chiang, W. 2006. A comparison of Maya and oracle bone scripts. *Visible Language*, 40(3): 310.

Da WEIH, D. 2023. Strategies of Graphing Ideas in Shang Oracle Bone Scripts. *The International Journal of Chinese Character Studies*, 6(2): 175–213.

Fu, X.; Yang, Z.; Zeng, Z.; Zhang, Y.; and Zhou, Q. 2022. Improvement of oracle bone inscription recognition accuracy: A deep learning perspective. *ISPRS International Journal of Geo-Information*, 11(1): 45.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Guan, H.; Yang, H.; Wang, X.; Han, S.; Liu, Y.; Jin, L.; Bai, X.; and Liu, Y. 2024. Deciphering Oracle Bone Language with Diffusion Models. *arXiv preprint arXiv:2406.00684*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, Z.; Cheung, Y.-m.; Zhang, Y.; Zhang, P.; and Tang, P.-l. 2024. Component-Level Oracle Bone Inscription Retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 647–656.

Li, B.; Dai, Q.; Gao, F.; Zhu, W.; Li, Q.; and Liu, Y. 2020. HWOBC-a handwriting oracle bone character recognition database. In *Journal of Physics: Conference Series*, volume 1651, 012050. IOP Publishing.

Li, J.; Chi, X.; Wang, Q.; Wang, D.; Huang, K.; Liu, Y.; and Liu, C.-l. 2024. A comprehensive survey of oracle character recognition: challenges, benchmarks, and beyond. *arXiv preprint arXiv:2411.11354*.

Liu, J.; Hong, Y.; and Zhang, X. 2009. *New Compilation of Oracle Bone Scripts*. Fujian People's Publishing House. ISBN 9787211058532.

Liu, S.; Zhou, Z.; Liu, Y.; Zhang, J.; and Qian, H. 2025. Language Representation Favored Zero-Shot Cross-Domain Cognitive Diagnosis. *arXiv preprint arXiv:2501.13943*.

Liu, Z.; and Liu, X. 2019. *Oracle Bone Script: Six Digit Numerical Code*. Sichuan Dictionary Publishing House. ISBN 9787557901509.

Mai, C.; Penava, P.; and Buettner, R. 2024. Oracle Bone Inscription Character Recognition based on a novel Convolutional Neural Network Architecture. *IEEE Access*.

Radford, A. 2018. Improving language understanding by generative pre-training.

Sheng, L.; Zhang, A.; Zhang, Y.; Chen, Y.; Wang, X.; and Chua, T.-S. 2024. Language models encode collaborative signals in recommendation.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, M.; and Deng, W. 2024. A dataset of oracle characters for benchmarking machine learning algorithms. *Scientific Data*, 11(1): 87.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, P.; Zhang, K.; Wang, X.; Han, S.; Liu, Y.; Jin, L.; Bai, X.; and Liu, Y. 2024b. Puzzle pieces picker: Deciphering ancient chinese characters with radical reconstruction. In *International Conference on Document Analysis and Recognition*, 169–187. Springer.

Wang, P.; Zhang, K.; Wang, X.; Han, S.; Liu, Y.; Wan, J.; Guan, H.; Kuang, Z.; Jin, L.; Bai, X.; et al. 2024c. An open dataset for oracle bone script recognition and decipherment. *arXiv preprint arXiv:2401.15365*.

Wang, S.; Guo, W.; Xu, Y.; Liu, D.; and Li, X. 2024d. Coarse-to-Fine Generative Model for Oracle Bone Inscriptions Inpainting. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, 107–114.

Weng, X.; Li, Y.; Hao, S.; and Hou, J. 2024. Oracle Bone Script Similiar Character Screening Approach Based on Simsiam Contrastive Learning and Supervised Learning. In *2024 2nd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV)*, 127–130. IEEE.

Wu, Z.; Su, Q.; Gu, K.; and Shi, X. 2024. A Cross-Font Image Retrieval Network for Recognizing Undeciphered Oracle Bone Inscriptions. *arXiv preprint arXiv:2409.06381*.

Yang, H.; Zhu, B.; Zhang, Y.; and Beiming, L. 2024. Study on Oracle Bone Image Preprocessing Based on Fusion of Denoising and Enhancement Techniques: Improving Text Detection Accuracy. In *2024 IEEE 7th International Conference on Electronic Information and Communication Technology (ICEICT)*, 452–456. IEEE.

Zhen, Q.; Wu, L.; and Liu, G. 2024. An Oracle Bone Inscriptions Detection Algorithm Based on Improved YOLOv8. *Algorithms*, 17(5): 174.