Specializing Large Models for Oracle Bone Script Interpretation via Component-Grounded Multimodal Knowledge Augmentation

Anonymous ACL submission

Abstract

Deciphering ancient Chinese Oracle Bone Script (OBS) is a challenging task that offers insights into the beliefs, systems, and culture of the ancient era. Existing approaches treat decipherment as a closed-set image recognition problem, which fails to bridge the "interpretation gap": while individual characters are often unique and rare, they are composed of a limited set of recurring, pictographic components that carry transferable semantic meanings. To leverage this structural logic, we propose an agent-driven Vision-Language Model (VLM) framework that integrates a VLM for precise visual grounding with an LLM-based agent to automate a reasoning chain of component identification, graph-based knowledge retrieval, and relationship inference for linguistically accurate interpretation. To support this, we also introduce OB-Radix, an expert-annotated dataset providing structural and semantic data absent from prior corpora, comprising 1,022 character images (934 unique characters) and 1,853 fine-grained component images across 478 distinct components with verified explanations. By evaluating our system across three benchmarks of different tasks, we demonstrate that our framework yields more detailed and precise decipherments compared to baseline methods.

1 Introduction

Oracle Bone Script (OBS), the earliest known mature writing system in China, holds significant historical and cultural value. Of the more than 4,500 identified OBS characters, only approximately one-third have been deciphered, leaving a vast corpus of glyphs in mystery (Li et al., 2024). Each undeciphered character represents a lost fragment of ancient institutions, technologies, and beliefs. However, the fragmented and stylised nature of OBS inscriptions, coupled with the requirement for profound paleographic and contextual expertise, renders manual decipherment exceptionally difficult.

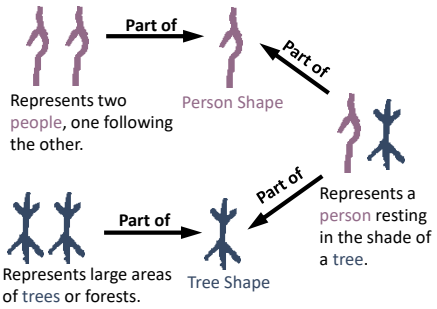

Figure 1: Oracle Bone Script (OBS), a pictographic writing system of semantic components.

In recent years, artificial intelligence has increasingly been leveraged for OBS interpretation (Fu et al., 2022; Wang et al., 2024a; Guan et al., 2024b; Jiang et al., 2023). Most existing methods treat decipherment as a closed-set image recognition task, largely neglecting the structural, semantic, and contextual nuances intrinsic to the script. This narrow focus results in information waste and introduces interpretive biases, as models lack the domain-specific knowledge required to generalise beyond known characters. While Vision-Language Models (VLMs) have demonstrated robust general image–text understanding (Liu et al., 2023; Caffagni et al., 2024), their capacity for fine-grained perception and expert reasoning remains a bottleneck. In low-resource, specialised domains like OBS, standard VLMs often suffer from "visual hallucinations" or a lack of linguistic depth, leading to bad performance (Chen et al., 2025; Ye, 2024).

Structurally, the OBS is an organised system of pictographic components, where each character is made up of discrete radicals that carry a distinct semantic weight and are frequently reused throughout the lexicon (Figure 1). This component-based architecture provides a critical logic bridge for decipherment: by identifying known pictographic components within an unknown glyph, we can systematically infer the meaning of characters that

are new to the model. To leverage this, we propose an Agentic Retrieval-Augmented Generation (Agentic RAG) framework that empowers VLMs with component-based semantic augmentation (Figure 2). To support this, we introduce OB-Radix, a new expert-annotated dataset comprising 1,022 Oracle character images (934 unique characters) and 1,853 fine-grained component images (478 distinct components), each paired with expert-verified semantic explanations. Finally, to evaluate whether our approach achieves expert-level capability, we design three progressively advanced benchmarks: (1) component-level retrieval, (2) component relationship inference, and (3) OBS interpretation generation. Experimental results demonstrate that our framework outperforms baseline methods, providing a more interpretable and linguistically accurate pathway for OBS decipherment.

In summary, our contributions are:

- We reformulate oracle bone script (OBS) interpretation as a *component-grounded, structure-aware reasoning task*, rather than a purely visual recognition problem, and instantiate this formulation with a multimodal framework that integrates component-level visual cues and graph-based retrieval.
- We construct OB-Radix, a component-level oracle bone script dataset, and build a knowledge graph that captures relationships among components, characters, and their semantic explanations, providing essential structured knowledge.
- We design comprehensive evaluations to assess both the accuracy and interpretability of our approach. Results show that our framework produces interpretations closely aligned with expert annotations and that the multi-agent extension offers enhanced semantic grounding.

2 Related Work

Deciphering of Oracle Bone Script. Existing research relies on a single image morphology model to explore AI reading paths. (Guan et al., 2024b) employs a diffusion approach to map oracle bone inscription images to modern Chinese characters, while (Qiao et al., 2024) leverages image generation to provide visual interpretive guidance. However, the former lacks integration of textual semantics, and the latter results in incomplete understanding due to the absence of textual guidance. Other studies applied diverse AI techniques (Fu et al., 2022; Jiang et al., 2023; Wang et al., 2024a; Gan et al., 2023) from different perspectives to aid in the decipherment of Oracle Bone Script.

Graph Retrieval-Augmented Generation for VLMs. Although large-scale VLMs demonstrate strong zero-shot generalization, they still exhibit noticeable performance drops when the underlying training corpora lack or misrepresent the necessary domain knowledge (Zhang et al., 2024; Minaee et al., 2024). To enhance the specialization of visual language approaches in particular domains, Retrieval-Augmented Generation (RAG) approaches are employed (Lin, 2024; Zhang et al., 2025). Unlike traditional fine-tuning, RAG dynamically retrieves relevant knowledge from external databases during inference, enabling VLMs to access domain-specific information on-demand without updating their pre-trained parameters. Additionally, to mitigate the potential noise present in general knowledge bases that may affect results, the concise representation provided by knowledge graphs are integrated, forming what is known as Graph RAG (Peng et al., 2024).

Oracle Bone Script Datasets. Most existing oracle bone script (OBS) datasets focus on *character-level* recognition, providing complete character images for end-to-end modeling, such as HUST-OBC (Wang et al., 2024b), EVOBC (Guan et al., 2024a), OBC306 (Huang et al., 2019), Oracle-50k (Han et al., 2020), and HWOBC (Li et al., 2020). While these datasets cover multiple historical scripts, they lack *component-level* annotations and therefore provide limited support for structural decomposition and interpretable semantic analysis. OracleFusion (Li et al., 2025a) introduces radical-level structures, bounding boxes, and semantic concepts for oracle characters, but its annotations remain region-based and lack expert-curated component entities, consistent semantic interpretations, and explicit inter-component relations, limiting its support for component-grounded reasoning and knowledge graph construction.

3 OB-Radix Dataset

We introduce OB-Radix, a dataset of hierarchical structural relations and grounded visual data meticulously curated by experts in paleography. Unlike prior character-level datasets (Wang et al., 2024b; Guan et al., 2024a; Huang et al., 2019; Han et al., 2020; Li et al., 2020) or those relying on ungrounded visual fragments (Hu et al., 2024) and text-only decompositions (Jiang et al., 2024), OB-

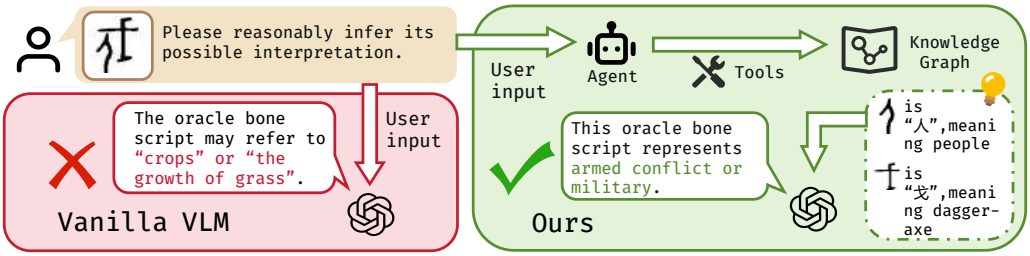

Figure 2: Comparison of our proposed framework and baselines. We design an agentic RAG framework to integrate component-level knowledge for structured semantic augmentation of OBS.

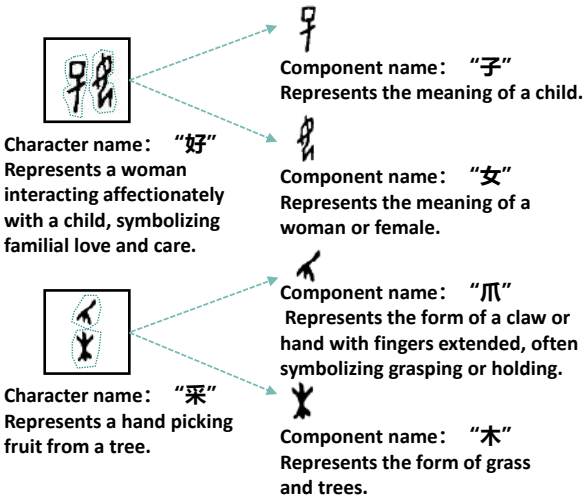

Figure 3: Our annotation of an oracle character at the component level.

Radix shifts the focus from the whole character to its constituent parts. This design enables models to learn and utilize the underlying compositional logic of the script, facilitating the interpretation of previously unseen or undeciphered glyphs through the identification of known pictographic elements. OB-Radix comprises 1,022 oracle character images covering 934 unique characters, along with 1,853 fine-grained component images spanning 478 distinct components.

Three archaeology doctoral students were tasked with identifying components based on their paleographic function and meaning, rather than relying solely on stroke continuity or visual salience. Specifically, annotators followed three core principles: (i) isolating components with distinct semantic roles, regardless of their visual scale; (ii) prioritizing semantic integrity over geometric completeness in cases of ambiguous boundaries; and (iii) maintaining uniform component labels across the corpus through a controlled vocabulary. These principles ensure that each component serves as a reliable semantic anchor, directly mapping visual regions to specific entries in our paleographic knowledge base for downstream reasoning.

Figure 3 showcases representative examples of this expert-level annotation, illustrating the decomposition of OBS characters into semantically meaningful components rather than arbitrary visual regions. To achieve high-precision segmentation, we utilized LabelMe (Russell et al., 2008) to perform semantic masking. Annotators were required to delineate the component region, after which the software automatically masked the largest contiguous region as the component body. To further refine the quality, experts manually adjusted key boundary points on each mask, ensuring the segmentation precisely captures the pictographic structure of the OBS. Given the specialized expertise required for such tasks, the curation process involved 70 total man-hours, representing a significant investment in high-fidelity data for the OBS domain. Further implementation details regarding the annotation tool, the expert workflow, and quality control measures are provided in Appendix A.1.

4 Method

As shown in Figure 4, our approach integrates visual analysis of OBS with structured knowledge reasoning through an agent-driven retrieval-augmented generation pipeline, comprising four parts: (1) a component identification module through character radicals retrieval as shown in Figure 4a, (2) an agent-driven knowledge graph retrieval module to dynamically query relevant entries as shown in Figure 4b, (3) a component relationship analysis and judgment module as shown in Figure 4c and (4) an interpretation generation module that integrates full character-level explanations as shown in Figure 4d. And the interpretation module supports two inference strategies: a VLM-based mode that directly fuses visual features with retrieved knowledge, and a multi-agent mode that

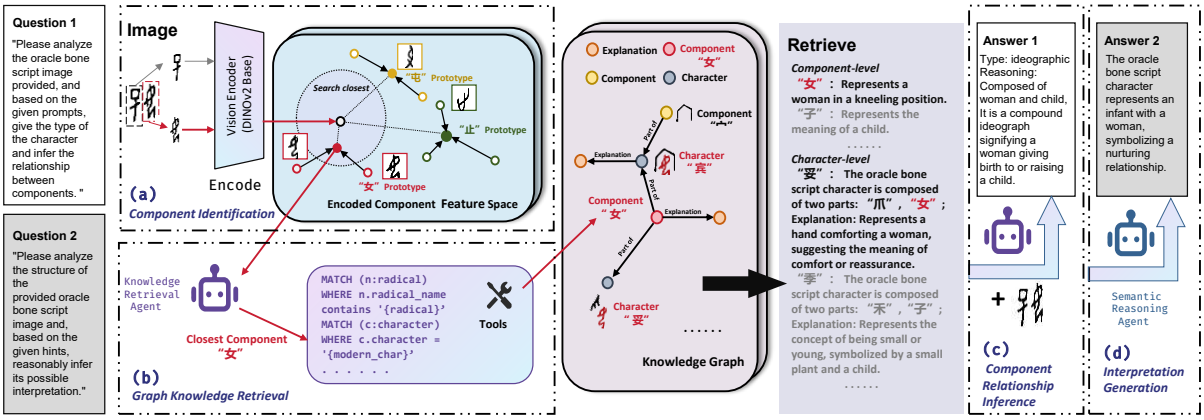

Figure 4: Detailed pipeline of our approach: (a) Component Identification Module identifies radical components from input OBS images; (b) Agent-Driven Graph Knowledge Retrieval retrieves relevant information from our constructed knowledge graph; (c) Component Relationship Inference uses VLMs to determine the structural relationships among components; (d) Interpretation Generation produces comprehensive semantic interpretations of oracle characters.

separates retrieval and reasoning into specialized agents, enhancing robustness and interpretability.

4.1 Component Identification

To identify radical components from input OBS images, we first utilize a Vision Transformer (ViT) architecture based on DINOv2 (Dosovitskiy et al., 2021; Oquab et al., 2024) to construct a component feature space, as it produces highly transferable features. Then, we adopt a prototype-based classifier following Prototypical Networks (Snell et al., 2017), as its class-level aggregation is well-suited to our low-data regime, improving robustness and reducing overfitting.

Specifically, Eq. (1) defines how the input radical image $\mathbf{x}$ is encoded by the DINOv2 encoder $f(\cdot)$ into a 768-dimensional vector $\mathbf{z}$, and we then compute its prototype $\mathbf{p}_c$ as the mean embedding of its support set $\mathcal{S}_c$ for each class c. Thus, given a query image $\mathbf{x}_q$, its embedding $\mathbf{z}_q = f(\mathbf{x}_q)$ is compared to all class prototypes using Euclidean distance $d(\cdot, \cdot)$, and then classified into the class with the nearest prototype.

$$\mathbf{z} = f(\mathbf{x}), \mathbf{z} \in \mathbb{R}^{768} \tag{1}$$

$$\hat{y} = \arg\min_c d(\mathbf{z}_q, \mathbf{p}_c) \tag{2}$$

Compared with directly using conventional classifiers or detectors, this design enables our model to make efficient use of limited labeled samples and enhances generalization in the low-resource setting of OBS component identification. An illustrative visualization of the component feature space construction is provided in Appendix A.2.

4.2 Agent-Orchestrated Graph Knowledge Retrieval

We construct a Knowledge Graph (KG) from OBRadix and character–component relations. For each test character, the PrototypeClassifier first predicts its most likely components; these predicted components are then used as *primary semantic cues* to query the KG. Rather than learning an unconstrained policy, we adopt a **cascading but largely fixed** retrieval pipeline, orchestrated by a tool-using LLM agent (Yao et al., 2023; Schick et al., 2023). The agent can call two external tools—*component explanation* and *characters-by-component*—and performs additional reasoning internally. Concretely:

- *Component-centric retrieval.* Given the predicted components, the agent first queries their explanations and searches for characters that contain these components, which typically provide the most direct semantic evidence.
- *Constrained synthesis.* When component-based retrieval yields weak or insufficient evidence, the agent internally performs variant lookup and modern–oracle mapping—without invoking external tools—to supplement the retrieved information. It then summarizes and reorders all evidence, both tool-obtained and internally inferred, into a concise, character-centric evidence bundle as input to the interpretation module.

To improve efficiency, we integrate a sim-

ple semantic-similarity cache following Jin et al. (2024), so that repeated or near-duplicate KG queries are served from cache. Overall, the agent acts as a lightweight orchestration layer over a deterministic retrieval cascade, ensuring that knowledge access is predictable and efficient while still providing rich, component-grounded context for downstream interpretation.

4.3 Component Relationship Inference

To move beyond black-box recognition, we design a module that leverages VLMs to infer the structural relationships among components. After the components are identified and the knowledge graph retrieval refines them, the system uses a VLM to jointly consider both visual embeddings and retrieved semantic information. The task requires the model to predict the inscription type of each oracle character, which can be categorized as ideographic, pictographic, or phono-semantic, and to generate a reasoning trace that explains how the components interact to form meaning. This process is illustrated in Figure 4, while the resulting output are presented in Figure 4c.

By conditioning the VLM on both structural and semantic cues, the module produces explanations that are not only accurate but also interpretable to human users. The component-level information is integrated into reasoning about character structure and provides the intermediate reasoning layer that connects recognition and interpretation generation.

4.4 Interpretation Generation

To generate full semantic interpretations of oracle characters, we design an inference pipeline that integrates visual recognition with knowledge-graph-based reasoning. Our framework supports two complementary modes of inference.

The first mode, *VLM Inference*, employs a VLM that jointly conditions on the visual embeddings of the inscription, component predictions from the PrototypeClassifier, and semantic prompts retrieved from the knowledge graph. By grounding ambiguous visual forms in curated historical evidence, the VLM produces interpretations that are semantically coherent and visually faithful.

Building upon this design, we further introduce a second mode, *Multi-Agent Inference*, inspired by recent advances in cooperative agent systems (Wu et al., 2023; Chang et al., 2024; Jin et al., 2025; Nguyen et al., 2025; Singh et al., 2025b; Wu et al., 2025). We use multi-agents to decouple retrieval and reasoning functions. A *Knowledge Retrieval Agent* plans and executes graph queries to gather relevant evidence, while a *Semantic Reasoning Agent* synthesizes this evidence with visual cues into structured, human-interpretable explanations. This separation improves robustness, reduces error propagation, and leverages the natural ability of large models to think after retrieval.

5 Experiments

To systematically evaluate whether our approach achieves expert-level capability in OBS interpretation, we design a series of experiments under expert guidance, structured around three progressively advanced tasks: (1) component-level retrieval as the foundation, (2) component relationship inference as the intermediate stage, and (3) OBS interpretation generation as the ultimate goal.

5.1 Metrics and baselines

We report ACC@k ($k \in \{1, 3, 5\}$) for component retrieval, and the accuracy of the oracle-character type classification for the component relationship inference experiment. We employ BERTScore-F1, MoverScore, ROUGE-1, and an LLM-as-a-Judge paradigm for OBS interpretation (Zhang* et al., 2020; Zhao et al., 2019; Lin, 2004; Zheng et al., 2023). To ensure evaluation impartiality (Li et al., 2025b), we instantiate the judge using Gemini 3 Flash (Team et al., 2023). Details of the LLM-as-a-Judge setup, including the evaluation rubric, prompting strategy, and the 0–1 scoring scale, are provided in Appendix A.4.

In the experimental tables, we use shorthand notations for VLMs. Specifically, *GPT* refers to GPT-5 (Singh et al., 2025a); *Claude* refers to Claude Opus 4.1 (20250805) (Anthropic, 2025); *GLM* refers to GLM-4.5V (V Team et al., 2025); and *Qwen* refers to Qwen3-VL-235B-A22B (Qwen Team, 2025).

5.2 Dataset Splitting

We adopted consistent dataset splitting strategies to ensure fair and realistic evaluation for all experiments. Specifically:

- **Component retrieval** (Section 5.3): Our OBRadix dataset, containing 478 distinct components, was divided into training and testing sets with a ratio of 7:3, respectively. Model performance was measured by Top-1, Top-3, and Top-5 accuracy.

Table 1: OBS component retrieval results.

Metric	ACC↑
Top-1	0.7795
Top-3	0.8855
Top-5	0.9157

- **Component relationship inference** (Section 5.4): We constructed a seen set of 528 annotated instances, each including both inscription type labels and expert-derived reasoning traces. Models were trained and evaluated on this split without data overlap, ensuring interpretability analysis was grounded in expert references.
- **Interpretation generation** (Section 5.5): To avoid leakage, our KG was built using 70% of the corpus, while the remaining 30% was held out for testing. This split applies to all experiments related to Section 5.5. It guarantees that characters used for evaluation had not appeared in training, thus presenting a realistic challenge of interpreting previously unseen instances.

5.3 Component Identification

The most essential prerequisite for understanding oracle bone characters lies in the ability to accurately recognize their constituent components, since these components serve as the fundamental units from which higher-level semantic and structural interpretations are derived. As summarized in Table 1, our approach achieves competitive recognition accuracy, demonstrating its effectiveness in capturing the visual and structural properties of OBS.

5.4 Component Relationship Inference

We evaluate whether VLMs capture the structural relationships among components, rather than treating OBS recognition as a black-box task. The task involves: (1) predicting the inscription type of a character (ideographic, pictographic, or phono-semantic), and (2) generating a textual explanation of component interactions. Representative examples comparing baseline and our enhanced pipeline are shown in Figure 5.

Table 2 reports classification and reasoning results. Our component-aware pipeline outperforms baselines across all metrics, confirming that explicit component-level knowledge improves both accuracy and interpretability. Qwen3-VL achieves the highest classification accuracy (0.599), while GPT-5 achieves the best performance under both BERTScore and LLM-as-a-Judge evaluation. Claude Opus 4.1 further shows the strongest fluency and alignment in reasoning (MoverScore, ROUGE-1).

Table 2: OBS component relationship inference results.

Category	Model	ACC↑	BERT↑	Mover↑	ROUGE-1↑	LLM-Judge↑
Baseline	GPT	0.364	0.497	0.310	0.007	0.237
	Claude	0.475	0.495	0.324	0.009	0.225
	GLM	0.447	0.519	0.293	0.010	0.113
	Qwen	0.350	0.503	0.318	0.012	0.165
Ours	GPT	$0.563^{+0.199}$	$\mathbf{0.670}^{+0.173}$	$0.472^{+0.161}$	$0.199^{+0.192}$	$\mathbf{0.435}^{+0.198}$
	Claude	$0.551^{+0.075}$	$0.648^{+0.152}$	$\mathbf{0.490}^{+0.166}$	$\mathbf{0.221}^{+0.212}$	$0.412^{+0.187}$
	GLM	$0.468^{+0.021}$	$0.606^{+0.088}$	$0.440^{+0.148}$	$0.139^{+0.129}$	$0.262^{+0.149}$
	Qwen	$\mathbf{0.599}^{+0.248}$	$0.658^{+0.156}$	$0.481^{+0.164}$	$0.212^{+0.200}$	$0.371^{+0.206}$

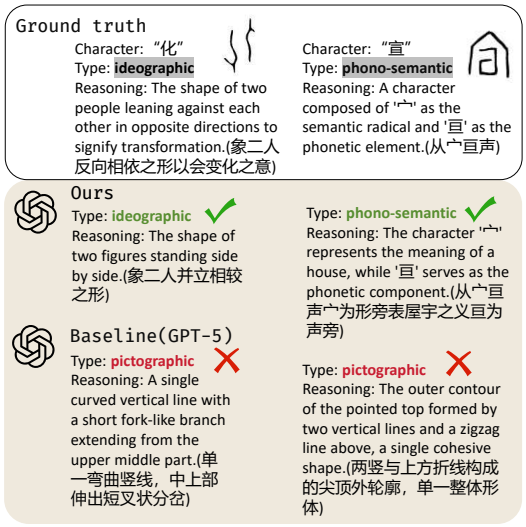

Figure 5: Reasoning examples for component relationship inference. *Ground truth* shows expert interpretations.

5.5 Interpretation Generation

This task provides a direct test of whether the system can go beyond recognition and structural reasoning to generate semantically meaningful interpretations.

We compare two categories of approaches: (1) *Baseline* models, where LLMs directly generate interpretations without access to the Knowledge Graph; and (2) *Agentic RAG* (ours), where the LLM retrieves supporting evidence from the graph before generating explanations. Performance was evaluated using BERTScore, MoverScore, ROUGE-1 and LLM-as-a-Judge with higher values indicating better alignment with expert-written ground truth. A concrete illustration is provided in A.3. Results are shown in Table 3.

The results clearly indicate the benefits of retrieval-augmented generation. Across all models,

Table 3: OBS interpretation generation results.

Category	Model	BERT↑	Mover↑	ROUGE-1↑	LLM-Judge↑
Baseline	GPT	0.633	0.393	0.227	0.102
	Claude	0.614	0.365	0.232	0.052
	GLM	0.634	0.338	0.275	0.042
	Qwen	0.636	0.362	0.264	0.076
Ours	GPT	**0.727**$^{+0.094}$	0.475$^{+0.082}$	0.321$^{+0.094}$	**0.558**$^{+0.456}$
	Claude	0.716$^{+0.102}$	0.474$^{+0.109}$	0.335$^{+0.103}$	0.382$^{+0.330}$
	GLM	0.706$^{+0.072}$	0.453$^{+0.115}$	0.337$^{+0.062}$	0.212$^{+0.170}$
	Qwen	0.722$^{+0.086}$	**0.471**$^{+0.109}$	**0.354**$^{+0.090}$	0.339$^{+0.263}$

Table 4: OBS relationship inference results.

Model	Retrieval	BERT↑	Mover↑	ROUGE-1↑	LLM-Judge↑
GPT	✓	**0.727**	**0.475**	0.321	**0.558**
		0.717$^{-0.010}$	0.469$^{-0.006}$	0.305$^{-0.016}$	0.542$^{-0.016}$
Claude	✓	0.716	0.474	0.335	0.382
		0.699$^{-0.017}$	0.451$^{-0.023}$	0.286$^{-0.049}$	0.372$^{-0.010}$
GLM	✓	0.706	0.453	**0.337**	0.212
		0.687$^{-0.019}$	0.415$^{-0.038}$	0.283$^{-0.054}$	0.180$^{-0.032}$
Qwen	✓	0.722	0.471	0.354	0.339
		0.711$^{-0.011}$	0.445$^{-0.026}$	0.326$^{-0.028}$	0.281$^{-0.058}$

the Agentic RAG pipeline consistently outperforms the baseline counterparts. For example, Qwen3-VL improves from 0.264 → 0.354 on ROUGE-1 and from 0.362 → 0.471 on MoverScore. Similarly, GPT-5 achieves the best BERTScore of 0.727 and the best LLM-as-a-Judge of 0.558 under the RAG setting, demonstrating stronger semantic alignment. These findings suggest that grounding interpretation generation in structured knowledge not only enhances factual accuracy but also produces outputs that are more coherent and interpretable.

5.6 Ablation Study

To isolate the contribution of the Agent-Driven Graph Knowledge Retrieval, we conducted an ablation experiment in which retrieval was disabled and only component category predictions were provided. The results are summarized in Table 4.

Across all four models, the absence of retrieval consistently reduces performance, confirming that the Oracle Knowledge Graph supplies non-trivial semantic context beyond visual recognition and component classification. Specifically, removing retrieval consistently degrades performance across all models and metrics. The performance drops are most evident on ROUGE-1 and LLM-as-a-Judge, indicating that Agent-Driven Graph Knowledge Retrieval provides crucial relational and contextual information beyond component category predictions. Moreover, the larger reductions in LLM-as-a-Judge compared to embedding-based metrics suggest that retrieval primarily improves higher-level semantic correctness rather than surface-level similarity. These results confirm that graph-based knowledge retrieval is essential for reliable OBS relationship inference.

Table 5: Performance comparison of multi-agent configuration.

Retriever	Reasoner	BERT↑	Mover↑	ROUGE-1↑	LLM-Judge↑
Qwen3-VL	DeepSeek-R1	**0.760**	**0.507**	**0.431**	0.531
	GPT-5	0.705	0.445	0.296	0.645
	Qwen3	0.734	0.470	0.361	0.494
GPT-5	DeepSeek-R1	0.733	0.476	0.402	0.413
	GPT-5	0.713	0.458	0.310	**0.657**
	Qwen3	0.729	0.454	0.366	0.464

5.7 Multi-Agent Collaboration

We further investigate a multi-agent setup, where the *Knowledge Retrieval Agent* (Retriever) first queries relevant entries from the Knowledge Graph, and the separate *Semantic Reasoning Agent* (Reasoner), instantiated with large language models such as GPT-5, DeepSeek-R1-250528 (Guo et al., 2025), or Qwen3-235B-A22B (Yang et al., 2025), subsequently composes the interpretation (Figure 4d). This separation is motivated by our earlier findings that factual grounding and reasoning fluency benefit from distinct model capabilities. As shown in Table 5, the multi-agent configurations generally outperform single-agent baselines across the evaluated metrics. We hypothesize that the Semantic Reasoning Agent is better equipped to process and integrate the textual information retrieved from the KG, leveraging its specialized capabilities for enhanced coherence and accuracy.

5.8 Human experts assessment study

To complement the above quantitative metrics, we conducted a human expert evaluation with two Ph.D. students in archaeology, using the 5-point Likert scale provided in A.5. For fairness, 10% of the held-out test set was selected, and participants were asked to evaluate the quality of generated interpretations along three pipelines: (1) the *Baseline pipeline* (direct generation using Qwen3-VL-235B-A22B), (2) the *RAG pipeline* (retrieval-augmented generation with Qwen3-VL-235B-A22B, and (3) the *Multi-Agent pipeline* (Qwen3-VL-235B-A22B as the Retrieval Agent and DeepSeek-R1-250528 as the Reasoning Agent).

Inter-rater reliability across all annotations was assessed using ICC3 (0.71) and Krippendorff's Al-

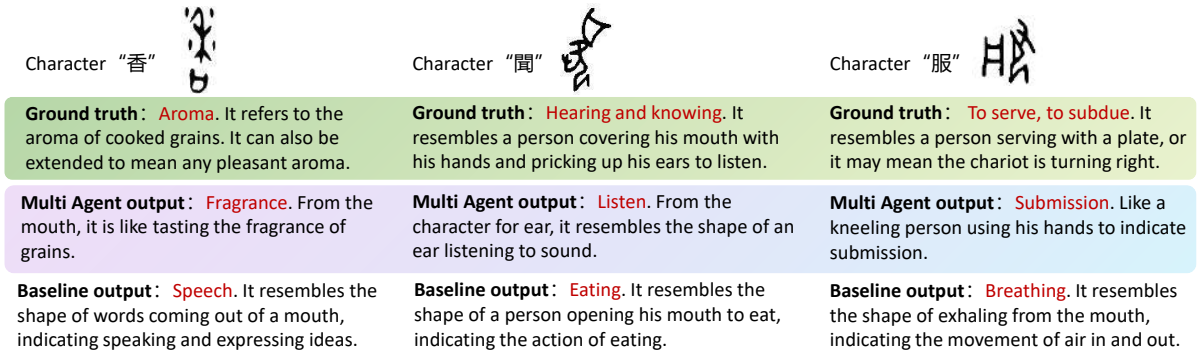

Figure 6: Comparison of approach outputs. *Character* displays the original Oracle bone characters; *Ground truth* provides the ground truth interpretations; *Multi-agent output* shows our multi-agent approach using Graph RAG; *Baseline output* presents results from the baseline approach.

Table 6: Results of interpretations conducted in English.

Category	Model	BERT↑	Mover↑	ROUGE-1↑	LLM-Judge↑
Baseline	GPT	0.152	-0.123	0.111	0.098
	Claude	0.136	-0.136	0.107	0.070
	GLM	0.036	-0.173	0.075	0.051
	Qwen	0.159	-0.126	0.117	0.068
Ours	GPT	$0.272^{+0.120}$	$0.034^{+0.157}$	$0.201^{+0.090}$	$\mathbf{0.797}^{+0.699}$
	Claude	$0.318^{+0.182}$	$0.071^{+0.207}$	$0.233^{+0.126}$	$0.752^{+0.682}$
	GLM	$0.318^{+0.282}$	$\mathbf{0.106}^{+0.279}$	$\mathbf{0.268}^{+0.193}$	$0.489^{+0.438}$
	Qwen	$\mathbf{0.320}^{+0.161}$	$0.075^{+0.201}$	$0.234^{+0.117}$	$0.619^{+0.551}$

pha (0.74), indicating substantial agreement between two PhD evaluators with expertise in archaeology. Average Likert scores on a five-point scale showed a clear performance hierarchy. The Multi-Agent Pipeline achieved the highest score of 3.433, followed by the KG-RAG pipeline at 2.133 and the baseline pipeline at 1.367. These human evaluation results are consistent with the automatic metrics and support the reliability of our experimental findings. To demonstrate the effectiveness of our multi-agent collaborative approach for oracle interpretation, we also qualitatively compare our approach with baseline methods in Figure 6.

6 Supplementary Experiments

In addition to the main experiments, we further conducted two supplementary studies to test the robustness and generalizability of our approach.

English Interpretation Generation. To investigate whether the models can generalize across languages, we constructed an English-version task, where the VLMs were required to output interpretations in English rather than Chinese. Results are reported in Table 6. Compared with the main Chinese results (Table 3), performance is notably lower across all metrics. This degradation is expected, since existing training corpora and retrieval databases are primarily constructed in Chinese, leading to weaker grounding in English. Nevertheless, the relative improvements of retrieval-augmented settings over baseline VLMs remain consistent, suggesting that our pipeline maintains cross-lingual robustness, albeit with a reduced ceiling. These results indicate the importance of developing parallel bilingual resources in paleographic studies to further support cross-linguistic generalization.

Variant Character Recognition. We evaluate a challenging variant character recognition setting, which requires models to associate visually distinct oracle character variants with a shared canonical form. Performance remains limited across all evaluated models, reflecting the intrinsic difficulty of this task in oracle bone script, where many variants lack explicit component or radical correspondences. Detailed results and expert-informed analysis are provided in Appendix A.6.

7 Conclusion

We propose a component-grounded framework for oracle bone script (OBS) interpretation that leverages the pictographic structure of the script and the relationships among its components. By integrating a component-structured Graph RAG with vision–language models, our approach supports interpretable OBS analysis. We further introduce a component-level oracle dataset and define three progressive tasks, including component retrieval, component relationship inference, and script interpretation, to enable structured evaluation. Experimental results demonstrate that knowledge graph augmentation improves both the accuracy and interpretability of OBS interpretation.

Limitations

In collaboration with paleographic experts, we identify several limitations of the current pipeline. Component recognition is not always precise or complete, and the system may occasionally introduce spurious elements. Moreover, a substantial portion of oracle characters still lack widely accepted interpretations, which inherently constrains the reliability of any automated analysis.

Future work may address these limitations by improving component recognition accuracy, expanding the coverage and quality of the underlying knowledge base, and extending the framework to better handle phono-semantic compounds, which remain challenging for current systems.

Finally, our current framework adopts a structured, retrieval-centric workflow rather than fully autonomous generation. This design limits flexibility and relies on external knowledge sources, reflecting the fact that existing VLMs lack intrinsic knowledge of Oracle Bone Script and may hallucinate under unconstrained generation. As base models evolve and acquire stronger domain understanding, future systems may reduce this dependency on explicit retrieval while maintaining philological reliability.

Ethical Considerations

This work uses publicly available Oracle Bone Script (OBS) resources and contains no personal, private, or sensitive data. All character- and component-level annotations were conducted by archaeology Ph.D. students with domain expertise in paleography, following authoritative references to ensure accuracy.

For human evaluation, two Ph.D. students participated voluntarily with informed consent. To reduce fatigue and ensure consistent evaluation conditions, only 10% of the held-out test set was assessed using a standardized Likert scale.

We acknowledge that automatic interpretation of cultural heritage materials may introduce errors or oversimplifications. Accordingly, the dataset, models, and experimental results presented in this work are intended solely as research aids to support scholarly analysis, and are not designed to replace expert judgment or authoritative paleographic interpretation.

References

Anthropic. 2025. System card addendum: Claude opus 4.1. Technical report, Anthropic.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand. Association for Computational Linguistics.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and 1 others. 2024. Main-rag: Multi-agent filtering retrieval-augmented generation. *arXiv preprint arXiv:2501.00332*.

Zijian Chen, tingzhu chen, Wenjun Zhang, and Guangtao Zhai. 2025. OBI-bench: Can LMMs aid in study of ancient script on oracle bones? In *The Thirteenth International Conference on Learning Representations*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Xuanming Fu, Zhengfeng Yang, Zhenbing Zeng, Yidan Zhang, and Qianting Zhou. 2022. Improvement of oracle bone inscription recognition accuracy: A deep learning perspective. *ISPRS International Journal of Geo-Information*, 11(1):45.

Ji Gan, Yuyan Chen, Bo Hu, Jiaxu Leng, Weiqiang Wang, and Xinbo Gao. 2023. Characters as graphs: Interpretable handwritten chinese character recognition via pyramid graph transformer. *Pattern Recognition*, 137:109317.

Haisu Guan, Jinpeng Wan, Yuliang Liu, Pengjie Wang, Kaile Zhang, Zhebin Kuang, Xinyu Wang, Xiang Bai, and Lianwen Jin. 2024a. An open dataset for the evolution of oracle bone characters: Evobc. *Preprint*, arXiv:2401.12467.

Haisu Guan, Huanxin Yang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. 2024b. Deciphering oracle bone language with diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15554–15567, Bangkok, Thailand. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu,

Yu Wu, Wu Z. F. Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 174 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638.

Wenhui Han, Xinlin Ren, Hangyu Lin, Yanwei Fu, and Xiangyang Xue. 2020. Self-supervised learning of orc-bert augmentator for recognizing few-shot oracle characters. In *Proceedings of the Asian Conference on Computer Vision*.

Zhikai Hu, Yiu-ming Cheung, Yonggang Zhang, Peiying Zhang, and Pui-ling Tang. 2024. Component-level oracle bone inscription retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 647–656.

Shuangping Huang, Haobin Wang, Yongge Liu, Xiaosong Shi, and Lianwen Jin. 2019. Obc306: A large-scale oracle bone character recognition dataset. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 681–688. IEEE.

Hanqi Jiang, Yi Pan, Junhao Chen, Zhengliang Liu, Yifan Zhou, Peng Shu, Yiwei Li, Huaqin Zhao, Stephen Mihm, Lewis C Howe, and 1 others. 2024. Oraclesage: Towards unified visual-linguistic understanding of oracle bone scripts through cross-modal knowledge fusion. *arXiv preprint arXiv:2411.17837*.

Runhua Jiang, Yongge Liu, Boyuan Zhang, Xu Chen, Deng Li, and Yahong Han. 2023. Oraclepoints: A hybrid neural representation for oracle character. In *Proceedings of the 31st ACM international conference on multimedia*, pages 7901–7911.

Chao Jin, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Xin Liu, Xuanzhe Liu, and Xin Jin. 2024. Ragcache: Efficient knowledge caching for retrieval-augmented generation. *arXiv preprint arXiv:2404.12457*.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. 2025. Disentangling memory and reasoning ability in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1681–1701, Vienna, Austria. Association for Computational Linguistics.

Bang Li, Qianwen Dai, Feng Gao, Weiye Zhu, Qiang Li, and Yongge Liu. 2020. Hwobc-a handwriting oracle bone character recognition database. In *Journal of Physics: Conference Series*, volume 1651, page 012050. IOP Publishing.

Bang Li, Donghao Luo, Yujie Liang, Jing Yang, Zengmao Ding, Xu Peng, Boyuan Jiang, Shengwei Han, Dan Sui, Peichao Qin, Pian Wu, Chaoyang Wang, Yun Qi, Taisong Jin, Chengjie Wang, Xiaoming Huang, Zhan Shu, Rongrong Ji, Yongge Liu, and Yunsheng Wu. 2024. Oracle bone inscriptions multimodal dataset. *Preprint*, arXiv:2407.03900.

Caoshuo Li, Zengmao Ding, Xiaobin Hu, Bang Li, Donghao Luo, AndyPian Wu, Chaoyang Wang, Chengjie Wang, Taisong Jin, Seven Shu, and 1 others. 2025a. Oraclefusion: Assisting the decipherment of oracle bone script with structurally constrained semantic typography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19893–19902.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025b. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Demiao Lin. 2024. Revolutionizing retrieval-augmented generation with enhanced pdf structure recognition. *arXiv preprint arXiv:2401.12599*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Thang Nguyen, Peter Chin, and Yu-Wing Tai. 2025. Ma-rag: Multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning. *arXiv preprint arXiv:2505.20096*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. Featured Certification.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.

Runqi Qiao, Lan Yang, Kaiyue Pang, and Honggang Zhang. 2024. Making visual sense of oracle bones for you and me. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12656–12665.

Qwen Team. 2025. Qwen3-VL: A more powerful large-scale vision-language model. https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef&from=research.latest-advancements-list. Accessed: 2025-05-28.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. 2008. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, and et al. Ananthram, Akhila. 2025a. Gpt-5 system card. Technical report, OpenAI.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025b. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30*, pages 4080–4090. Curran Associates, Inc.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. 2024a. Puzzle pieces picker: Deciphering ancient chinese characters with radical reconstruction. In *International Conference on Document Analysis and Recognition*, pages 169–187. Springer.

Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Jinpeng Wan, Haisu Guan, Zhebin Kuang, Lianwen Jin, Xiang Bai, and 1 others. 2024b. An open dataset for oracle bone character recognition and decipherment. *Scientific Data*, 11(1):976.

Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. 2025. Agentic reasoning: A streamlined framework for enhancing LLM reasoning with agentic tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28489–28503, Vienna, Austria. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 3(4).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Cheng Ye. 2024. Exploring a learning-to-rank approach to enhance the retrieval augmented generation (rag)-based electronic medical records search engines. *Informatics and Health*, 1(2):93–99.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.

Ruixiang Zhang, Yu Wang, Weiyang Yang, Jun Wen, Weizhi Liu, Shipeng Zhi, Guangzhou Li, Nan Chai, Jiaqi Huang, Yongyao Xie, and 1 others. 2025. Plantgpt: An arabidopsis-based intelligent agent that answers questions about plant functional genomics. *Advanced Science*, page e03926.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Appendix

A.1 More details on dataset construction

To ensure fine-grained component-level annotation, we adopted **LabelMe**[1] as the primary tool for manual segmentation of Oracle Bone Script images. LabelMe allows annotators to draw polygonal masks directly on images, making it well suited for the irregular shapes and complex outlines of Oracle characters, as shown in Figure 7.

Each annotation task was conducted by archaeology PhD students who followed authoritative decipherment references. Annotators were compensated for their annotation efforts, with a total payment of 2,450 RMB across all tasks. The process began with drawing precise polygons around each component within a character image. These polygons were then exported into JSON format, which stores the coordinates of the segmentation boundaries together with the corresponding component labels. To improve annotation consistency, we designed a standardized guideline specifying:

- **Segmentation granularity:** ensuring that even small components with distinct semantic functions were delineated separately.
- **Boundary precision:** refining polygon edges to closely follow character contours, especially in cases where strokes overlapped or eroded.
- **Label consistency:** using controlled vocabularies for component names to avoid ambiguity across annotators.

As illustrated in Figure 7, the annotation workflow produces both the original oracle character and its corresponding component-level masks, which are paired with expert-verified semantic explanations. To ensure annotation reliability, all annotations were performed by archaeology PhD students following authoritative decipherment references, and were subsequently cross-checked to resolve ambiguous boundaries and label inconsistencies.

This expert-curated procedure ensures that **OBRadix** achieves high annotation quality and interpretive reliability, laying the foundation for downstream tasks in component recognition and semantic inference.

A.2 Illustration of Component Feature Space Construction

Figure 8 provides an intuitive illustration of the component feature space construction process described in Section 4.1. Each radical image is first encoded by the DINOv2 encoder into a high-dimensional embedding vector. Images belonging to the same component class form compact clusters in the embedding space, while prototypes (class means) serve as representative anchors for classification.

As shown in the figure, query samples are classified based on their distances to class prototypes rather than individual training instances. This geometry encourages intra-class compactness and inter-class separability, which is particularly beneficial in low-resource scenarios where only a few labeled examples are available per component.

Such a structured feature space allows the model to generalize effectively to unseen samples while mitigating overfitting, making prototype-based classification well suited for OBS component identification.

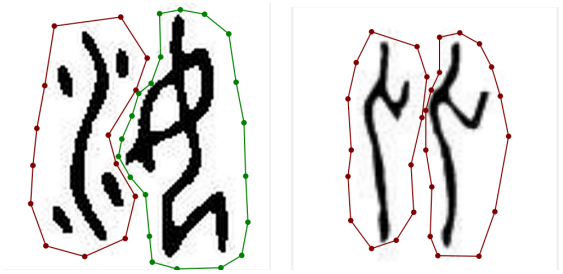

Figure 7: Two images of oracle bone characters segmented by LabelMe.

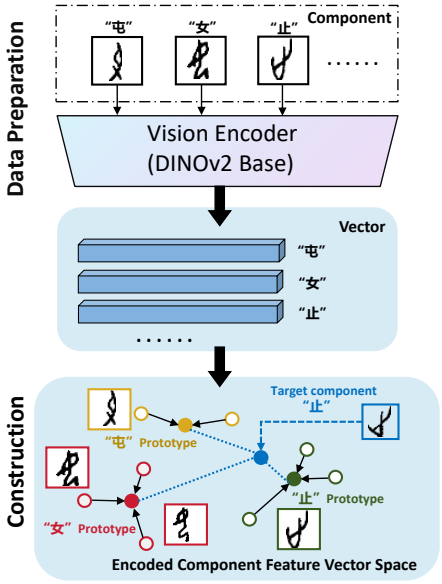

Figure 8: Construction of Vector Space.

[1] https://github.com/wkentaro/labelme

A.3 Oracle Bone Script Interpretation Example

This section provides an example of oracle bone script (OBS) interpretation generated by our models to illustrate the difference between the *Baseline* and *Agentic RAG* approaches, as shown in Figure 9.

A.4 LLM-as-a-Judge Evaluation

To evaluate the semantic correctness of OBS interpretation generation, we adopt an LLM-as-a-Judge evaluation paradigm (Zheng et al., 2023). In this setting, a large language model is prompted to compare a model-generated interpretation with a reference interpretation provided by domain experts and to assign a scalar score in the range of 0 to 1, where higher scores indicate better semantic alignment.

The evaluation focuses on semantic consistency rather than surface-level lexical overlap, taking into account the correctness of key entities and participants, core events and relations, semantic modifiers, as well as potential hallucinations or critical omissions. We instantiate the judge with Gemini 3 Flash using a fixed prompt template and temperature set to zero to ensure deterministic behavior.

Prompt Template. The exact prompt used for the LLM-as-a-Judge evaluation is shown below:

System Instruction:
You are a rigorous semantic assessment expert. You are responsible for scoring the semantic consistency of the sentence to be scored based on the reference sentence.
Scoring Criteria (0.00–1.00): Output a score rounded to the nearest 0.01 (e.g., 0.66, 0.92).

- `0.80–1.00 (Perfect)`: Semantically equivalent. Core information and details are accurate, with only reasonable paraphrasing.
- `0.60–0.79 (Excellent)`: Core semantics are accurate. Minor modifiers may be missing, but there are no factual errors.
- `0.40–0.59 (Acceptable)`: Contains key information, but omits some important details or contains minor ambiguities.
- `0.20–0.39 (Poor)`: Significant omission of key information or inclusion of obvious hallucinations, leading to semantic distortion.
- `0.00–0.19 (Failure)`: Completely irrelevant, opposite meaning, or nonsense.

User Instruction: Please answer in this format:
`Score: [a number between 0.00 and 1.00]`

A.5 Oracle Bone Script Interpretation Questionnaire

The questionnaire consists of 30 candidate interpretations of oracle bone script characters. Specifically, we curated 10 distinct characters, each of which is associated with three alternative interpretations reflecting different reasoning pipelines. To avoid introducing bias from fixed presentation sequences, the three interpretations corresponding to the same character were randomly permuted prior to distribution. This randomization was applied independently across pipelines, ensuring that participants evaluated the interpretations without being influenced by a consistent order effect. Consequently, the design of the questionnaire facilitates a more balanced and reliable assessment of the comparative quality of the proposed interpretation methods (see Figure 10).

Table 7: Variant character search (39 samples).

Model	Top-1@ACC	Top-5@ACC	Top-10@ACC
GPT-5	5.13% — 2	5.13% — 2	5.13% — 2
Claude Opus 4.1	0.00% — 0	2.56% — 1	5.13% — 2
GLM-4.5V	2.56% — 1	2.56% — 1	2.56% — 1
Qwen3-VL-235B-A22B	2.56% — 1	2.56% — 1	5.13% — 2

A.6 Variant Character Recognition

Oracle Bone Script contains a large number of variant character forms due to its extended historical usage and diverse engraving practices. Paleographic studies indicate that the script was used for nearly two centuries by multiple engraving groups, resulting in substantial visual variation across characters, often without shared radicals or explicit component structures. To study this phenomenon, we curated 39 variant character pairs and evaluated whether models could associate each variant with its canonical form. As shown in Table 7, recognition accuracy remains low across all evaluated models, with no method achieving strong Top-1 or Top-10 performance, reflecting the intrinsic difficulty of this task.

This difficulty arises because many variants lack consistent substructures that can be captured by component- or radical-based visual representations, making compositional cues insufficient in this setting. These results suggest that effective variant recognition may require targeted supervision, explicit variant–canonical mappings, or deeper integration of expert paleographic knowledge, which we leave for future work.

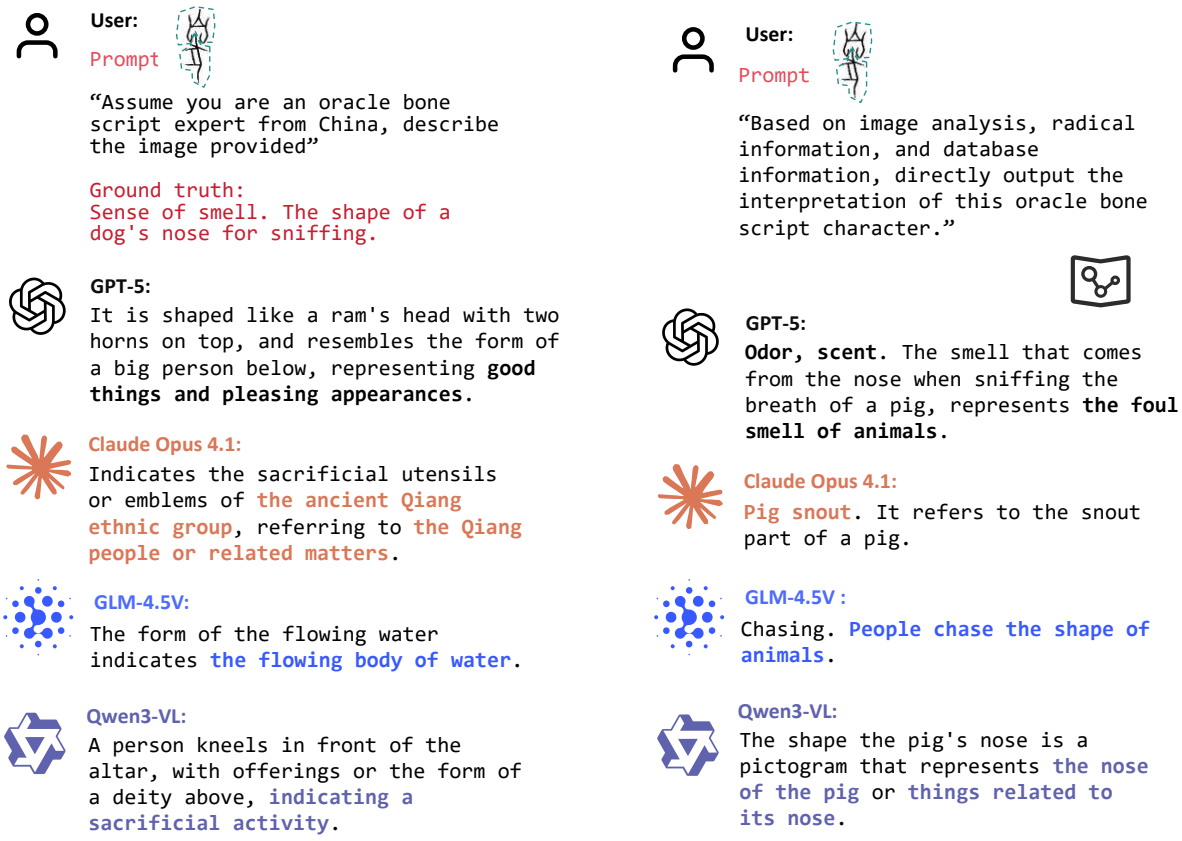

Figure 9: The left side shows the baseline outputs, while the right side shows our results.

Instructions

Do you agree with this interpretation of the oracle bone script? (5-point scale) Please tick (√) the score that best reflects your agreement with each oracle-bone-script interpretation below.

Scoring Scale

- **(5) Completely Agree**: The interpretation fully matches the oracle bone script's glyph original meaning and construction logic, without any semantic distortion or historical deviation.
- **(4)Basically Agree**: The core interpretation is correct (matches the glyph original meaning and mainstream views), but there are extremely minor expression flaws (such as imprecise terminology) or omissions of secondary information (such as not mentioning rare usages), which do not affect the overall accuracy of the interpretation.
- **(3) Neutral**: The interpretation has "ambiguity" or "points of controversy" there is no clear evidence to prove it wrong, nor does it fully match authoritative interpretations; possibly due to the oracle bone script's own glyph defects, ongoing academic debates, or the interpretation only covering partial possibilities.
- **(2) Basically Disagree**: The core interpretation is wrong (violates the glyph original meaning or mainstream academic views), but there are a few reasonable elements (such as correct partial glyph disassembly, or involving secondary usages of the character); the overall interpretation deviates from the essence, but not completely baseless.
- **(1) Completely Disagree**: The interpretation completely contradicts the oracle bone script's glyph, construction logic, and academic consensus; unrelated to any known usage of the character.

Example

Output	Score				
	1	2	3	4	5
• It is like placing something on a stand with two hands. Four hands hold the object and place it on the ground or on a stand, which means to place or put it.	○	○	○	○	○
• The image of holding jade in both hands and offering it to the altar represents a sacrificial ceremony.	○	○	○	○	○
• It resembles four hands holding up a tube.	○	○	○	○	○

Figure 10: Questionnaire