OracleAnalyser: Analyzing Implicit Semantics of Oracle Bone Scripts through MLLMs with Post-training

Anonymous ACL submission

Abstract

With the advancement of artificial intelligence, research on oracle bone scripts has entered a new era. However, existing methods and benchmarks remain largely confined to recognition tasks, overlooking the equally crucial aspect of oracle bone analysis. To address this gap, we propose OracleAnalyser, a reasoning framework for oracle bone analysis based on post-training techniques. Specifically, we fine-tune Qwen2.5-VL-3B-Instruct through multiple post-training stages and introduce a new preference optimization algorithm, Stable Focal Preference Optimization (SFPO), tailored to the characteristics of oracle bone datasets. In addition, we release both an oracle bone reasoning dataset and an oracle bone preference dataset, and further construct a new benchmark to evaluate models' analytical capabilities for oracle bone scripts. Extensive experiments validate the superior analytical performance of OracleAnalyser, which achieves remarkable results with only 3B parameters, surpassing models with substantially larger scales.

1 Introduction

Oracle Bone Script (OBS) represents the earliest known mature writing system discovered in China, originating in the Shang dynasty. They were primarily engraved on turtle plastrons and animal bones for divinatory purposes. Rooted in pictographic forms, these characters not only documented aspects of contemporary society and daily life, but also carried invaluable historical significance, offering a unique window into the language and culture of early Chinese civilization. To date, more than 4500 oracle bone scripts have been identified, yet only about 1600 scripts have been deciphered and linked to modern Chinese characters, leaving the majority still undeciphered (Wang and Deng, 2024). Analyzing and interpreting these undeciphered characters is crucial for the preservation of cultural heritage and carries immense value for archaeology, historical research, and philology. Consequently, oracle bone analysis has become an important and active research topic in the academic community (Jiang et al., 2023; Peng et al., 2025).

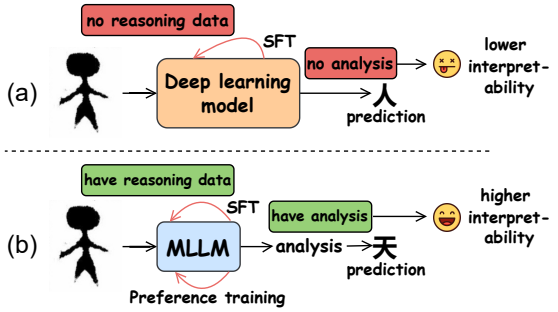

Figure 1: The differences between the previous deep learning approaches and ours. (a) Previous approaches can not perform analytical reasoning, resulting in limited interpretability. (b) our method can generate analysis before recognition, making it more conducive to the decipherment of oracle bone scripts.

In order to decipher these ancient scripts, recent studies have begun applying modern AI techniques. For instance, OBSD (Guan et al., 2024b) employs a conditional diffusion model to learn the patterns of font evolution for oracle bone script decipherment, while Puzzle Pieces Picker (Wang et al., 2024b) utilizes Transformers to reconstruct strokes and decode these enigmatic characters. However, most existing works focus primarily on recognition, often neglecting deeper analysis of the script. Moreover, the majority of current benchmarks (Chen et al., 2024; Wang et al., 2024a) are designed for recognition tasks and lack evaluation metrics for analytical capabilities. We argue that the ability of recognizing oracle bone script resembles memorization, whereas analytical competence is more crucial for effective decipherment.

Post-training techniques combined with reasoning have led to significant breakthroughs in general domains (Tie et al., 2025). For instance, DeepSeek-R1 (Guo et al., 2025) leverages large-scale rein-

forcement learning to elicit remarkable reasoning capabilities, demonstrating its effectiveness across various benchmarks. MCTS-DPO (Xie et al., 2024) employs Monte Carlo Tree Search (MCTS) to iteratively collect preference data and applies Direct Preference Optimization (DPO) (Rafailov et al., 2023) to update the LLM policy. In the multimodal domain, InternVL3.5 (Wang et al., 2025c) introduces a Cascade Reinforcement Learning (Cascade RL) framework to advance versatility, reasoning capability, and inference efficiency. Similarly, Zhang et al. (Zhang et al., 2024) adopt a two-stage post-training strategy, combining Supervised Fine-tuning (SFT) and DPO, to enhance the chain-of-thought reasoning ability of VLMs.

These achievements inspire us to consider: if the analysis of oracle bone characters can be regarded as a specific form of reasoning, could such post-training approaches be leveraged to facilitate their interpretation and decipherment? Motivated by this idea, we propose OracleAnalyser, a reasoning framework for oracle bone analysis based on post-training techniques. Specifically, we construct a new oracle bone reasoning dataset and use it to fine-tune Qwen2.5-VL-3B-Instruct (Bai et al., 2025) under supervised learning, obtaining OracleAnalyser-sft. We then employ OracleAnalyser-sft for inference to generate positive and negative sample pairs, thereby building a preference dataset tailored for oracle bone analysis. Furthermore, we propose a new algorithm for oracle bone analysis, Stable Focal Preference Optimization (SFPO), to further train OracleAnalyser-sft, resulting in the final OracleAnalyser. The differences between our method and previous deep learning approaches are illustrated in Figure 1. Meanwhile, we design a new benchmark to evaluate the intermediate analytical reasoning output of MLLMs rather than focusing only on recognition outcomes.

Our contributions can be summarized as follows: (1) We release a newly constructed oracle bone reasoning dataset and the first oracle bone preference dataset for preference training. In addition, a new benchmark is designed for evaluating oracle bone analysis. (2) We propose the oracle bone analysis framework named OracleAnalyser, and introduce a new preference optimization objective, SFPO, tailored to the characteristics of the oracle bone dataset, establishing the first work that integrates reasoning with multiple post-training for oracle bone analysis. (3) We conduct extensive experiments and demonstrate that OracleAnalyser significantly outperforms other MLLMs in oracle bone analysis. Moreover, we rigorously verify the roles of different components through ablation studies.

2 Related works

Intelligent recognition and decipherment of oracle-bone scripts. For datasets, EVOBC compiles multi-era glyphs for evolution-aware decipherment (Guan et al., 2024a), and HUST-OBC provides a large-scale corpus and benchmark (Wang et al., 2024a). OBI-Bench evaluates LMMs end to end on oracle-bone tasks, exposing gaps in fine-grained perception and domain knowledge (Chen et al., 2024). For methods, P^3 reconstructs radicals to turn undeciphered characters into analyzable sequences (Wang et al., 2024b), and OracleSage employs the MLLM and fuses hierarchical visual structure with knowledge-graph reasoning (Jiang et al., 2024). CFIRN casts decipherment as cross-font or cross-era retrieval with a Siamese, multi-scale model (Wu et al., 2024), and OBSD utilizes a conditional diffusion-based strategy for generating vital clues for decipherment (Guan et al., 2024b).

The reasoning progress of MLLMs. Recently, reinforcement fine-tuning (RFT) has significantly advanced the reasoning capabilities of MLLMs (Wang et al., 2025a; Li et al., 2024; Liu et al., 2025c; Chen et al., 2025b; Wang et al., 2025b). Sun et al. systematically review the theoretical foundations and practical progress of multimodal-RFT(Sun et al., 2025). Li et al. propose CLS-RL and No-Thinking-RL, and argue that explicit reasoning is not always beneficial (Li et al., 2025). At scale, Kimi-VL integrates Mixture-of-Experts (MoE), long-context modeling, and long-CoT supervision to enhance multimodal reasoning (Team et al., 2025b,a). InternVL-3.5 adopts a cascaded reinforcement learning framework, reporting substantial gains across diverse multimodal reasoning benchmarks (Wang et al., 2025c). Along the R1-style research line, VLM-R1 adapts the R1 paradigm to vision-language models, achieving better generalization than standard SFT (Shen et al., 2025). Similarly, Perception-R1 demonstrates perception-policy improvements on COCO and related benchmarks (Yu et al., 2025), VisRL directly optimizes the intention-attention-reasoning loop via reinforcement learning (Chen et al., 2025a), and OpenVLThinker employs iterative SFT-RL to transfer text-based reasoning paradigms into multimodal contexts (Deng et al., 2025).

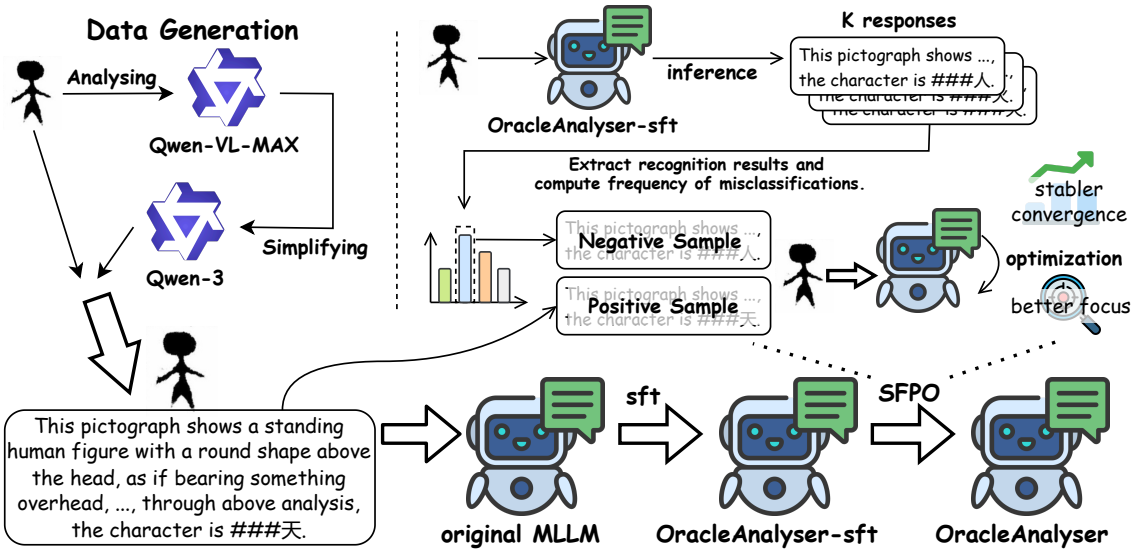

Figure 2: The overall framework of OracleAnalyser. It employs reasoning combined with post-training techniques to analyze and recognize oracle bone scripts. The original MLLM is Qwen2.5-VL-3B-Instruct. The post-training process consists of supervised fine-tuning (SFT) and the proposed SFPO. Compared with DPO, SFPO achieves stabler convergence and provides better focus on valuable samples.

3 Method

3.1 Overview of Framework

Robust oracle bone analyzing capabilities are crucial for deciphering these enigmatic symbols (Qiao et al., 2025). Motivated by the demonstrated effectiveness of combining reasoning with post-training techniques in general domains, we propose OracleAnalyser, a reasoning framework for oracle bone analysis based on post-training methods. The overall framework is illustrated in Figure 2. Specifically, we first prompt the powerful Qwen models to construct an oracle bone reasoning dataset, which is then fed into Qwen2.5-VL-3B-Instruct. By applying an autoregressive loss for supervised learning, we obtain OracleAnalyser-sft. Next, images from the oracle bone reasoning dataset are input into OracleAnalyser-sft for multiple inference rounds to generate positive and negative sample pairs, forming the oracle bone preference dataset. In addition, to address the limitations of direct preference optimization(DPO) in oracle bone analysis area, we introduce a new optimization algorithm, SFPO, and apply it to further preference training of OracleAnalyser-sft on the newly constructed dataset. We will elaborate on the construction process of the oracle bone reasoning dataset in Section 3.2, present the supervised fine-tuning (SFT) procedure in Section 3.3, and describe the details of preference data construction as well as the operation of SFPO in Section 3.4.

3.2 Construction of Oracle Bone Reasoning Dataset

We select the OBC subset of the Evolution Dataset of Oracle Bone Characters (EVOBC) as the data source, where each entry contains an image of an oracle bone character and its corresponding modern Chinese character. However, since this dataset includes certain archaic characters that are no longer in use, we filter it based on the list of 3500 commonly used Chinese characters, resulting in a collection of 2807 characters. As multiple oracle bone images may correspond to the same Chinese character, to avoid distributional imbalance, we restrict the maximum number of oracle bone images per character to 10, yielding approximately 6000 samples in total. From this collection, we retain 5500 samples, of which 5000 are randomly selected for the training set and 500 for the in-domain test set, while the remaining 500 samples are reserved as an out-of-domain test set.

We input oracle bone images into Qwen-VL-MAX (Bai et al., 2025) with the following prompt: `This is the oracle bone of {hz}. Describe the script from a pictographic perspective, but do not mention {hz} explicitly in your description.` Here, `{hz}` denotes the modern Chinese character corresponding to the oracle bone image. With knowledge of the associated modern character, Qwen-VL-MAX can provide a more accurate analysis of the oracle

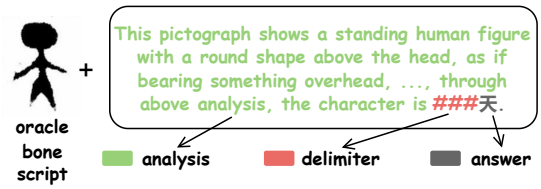

Figure 3: The format of one sample in oracle bone reasoning dataset.

bone symbol. Subsequently, the analysis generated by Qwen-VL-MAX is refined using Qwen-3 (Yang et al., 2025) to remove redundant information, guided by the prompt: `For the analysis of an oracle bone script{analysis}, please help me make it more concise.` Here, `{analysis}` refers to the preliminary analysis corresponding to the oracle bone image. Finally, we perform manual verification and post-processing on the oracle bone images and their corresponding analyses. Each data entry therefore consists of an oracle bone image, its analysis, and the corresponding modern Chinese character. The specific format of one sample is shown in Figure 3.

3.3 Supervised Fine-tuning Process

We adopt Supervised Fine-tuning (SFT) as the first stage of post-training. In this stage, the model could gradually captures the intrinsic visual patterns of oracle bone scripts, and develop fundamental reasoning abilities for analyzing oracle bone characters. Moreover, this stage provides a stable initialization for subsequent preference optimization, serving as a critical bridge between raw pretrained knowledge and the specialized reasoning abilities required for oracle bone analysis.

Specifically, we employ Qwen2.5-VL-3B-Instruct as backbone, and the image of oracle bone character, together with the prompt: `This is an oracle bone character. Please try to analyze it and identify it, and use ### to separate the analysis from the recognition result`, are fed into the MLLM to generate structured analyses of the symbols. During training, the model's output is compared against the ground-truth annotations, and the autoregressive loss is introduced for optimization. The formula is shown as follows:

$$\mathcal{L}_{AR} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t},\, I,\, P; \theta)\,, \qquad (1)$$

where I denotes the visual embedding obtained from the image input through the vision encoder, and P represents the prompt token. After concatenating the visual embedding and the prompt token, they are used as input to the decoder. y_t is the t-th token in the output sequence. T represents the sequence length, and θ denotes the learnable parameters of the model.

After SFT, we obtain the trained model, OracleAnalyser-sft. This process not only equips the model with basic recognition and interpretation skills, but also ensures that it captures the semantic correspondence between oracle bone characters and their modern counterparts, which could avoid converging to trivial or unstable solutions when trained in the preference optimization.

3.4 Stable Focal Preference Optimization

As one of the mainstream post-training techniques, preference learning can optimize the generation quality of models in the absence of explicit labeled data by leveraging relative preference information. This not only helps MLLMs capture fine-grained semantic differences during training, but also enhances the stability and diversity of their outputs. In the domain of oracle bone analysis, preference learning can further refine the model's analytical capability on top of OracleAnalyser-sft.

The quality of the preference dataset directly determines the effectiveness of preference training; thus, constructing high-quality positive and negative sample pairs is crucial. Specifically, we input the oracle bone images and prompts from the oracle bone reasoning dataset into OracleAnalyser-sft, where the model performs five inference traces for each image. The recognition results are then extracted from these outputs through delimiters, and we record the frequency of each erroneous recognition. Formally, we obtain $\{\mathbf{s}_i, |\mathbf{s}_i|\}_i$, where $\mathbf{s}_i$ denotes the set of reasoning processes yielding the same incorrect recognition result, and $|\mathbf{s}_i|$ is the number of samples in that set. From the set $\mathbf{s}_i$ with the highest $|\mathbf{s}_i|$, we randomly select one reasoning process as the negative sample, while the ground truth serves as the positive sample, together forming the preference pair for the corresponding oracle bone image. This procedure is repeated for each oracle bone image, resulting in over 4700 preference pairs. Subsequently, we manually filter these preference pairs by removing those that are difficult to distinguish, resulting in over 3400 preference pairs that constitute the oracle bone preference dataset.

By constructing the dataset, we preferentially select negative samples aligning with the model's

own cognitive biases, while filtering out trivial cases, thereby facilitating effective optimization of the model's analysis capability.

Direct Preference Optimization (DPO) is a widely adopted algorithm in preference learning. Let $\mathcal{D} = (x^{(i)}, y_w^{(i)}, y_l^{(i)})_{i=1}^{N}$ represent the oracle bone preference dataset, where $x^{(i)} \in \mathcal{X}$ consists of a prompt and the corresponding oracle bone image. Each $(y_w^{(i)}, y_l^{(i)})$ denotes a pair of candidate answers, with the preference of $y_w^{(i)} \succ y_l^{(i)}$. Following the Bradley–Terry model, the preference distribution p^* can be formulated as:

$$p^*(y_w^{(i)} \succ y_l^{(i)} \mid x^{(i)}) = \sigma\big(r^*(x^{(i)}, y_w^{(i)}) - r^*(x^{(i)}, y_l^{(i)})\big), \tag{2}$$

where σ is the sigmoid function and $r^*(x, y)$ is the latent reward model used to generate the ground-truth preference. To model the preference, DPO uses a reparametrization trick to express it in terms of the optimal policy π^*:

$$r^*(x, y) = \beta \log \frac{\pi^*(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x), \tag{3}$$

where $Z(x)$ is the partition function only based on x. Through equation 3, the maximum likelihood estimation objective of DPO is as follows:

$$\begin{aligned}\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log p(y_w \succ y_l \mid x)\right] \\ &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\log \sigma\Bigg(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} \\ &\quad - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\Bigg)\Bigg],\end{aligned} \tag{4}$$

where minimizing $\mathcal{L}_{\text{DPO}}$ encourages π_θ to align with π^*, thereby increasing the likelihood of producing responses with higher quality.

However, in the domain of oracle bone analysis, directly applying DPO can lead to problems. Specifically, the positive samples in the oracle bone preference dataset still contain noise, while some negative samples include partially correct analyses. This may cause the policy model to overfit these imperfect preferences and deviate excessively from the reference model, thereby undermining the theoretical foundation of DPO, distorting the optimization objective, and potentially leading to over-optimization or reward hacking. In addition, DPO tends to emphasize learning from response pairs that the model struggles to rank. Yet, the oracle bone preference dataset still contains a number of such difficult preference pairs, in which both positive and negative samples may include partially correct or reasonable content but are labeled with a forced preference. Thus, DPO training may exaggerate minor differences, leading the model to over-interpret or fabricate distinguishing features.

To address the aforementioned issues, we propose a new preference optimization algorithm, Stable Focal Preference Optimization (SFPO), building upon DPO. To prevent over-optimization in the later stages of training and stabilize model convergence, it is necessary to ensure that the theoretical assumptions of DPO approximately hold, namely, that the divergence between the policy model and the reference model remains within a reasonable range (Liu et al., 2025a). At the same time, to avoid over-interpretation, we aim to reduce the emphasis on learning from response pairs that the model struggles to rank and instead focus on preference pairs with reliable implicit reward estimates. Accordingly, the loss function of the newly defined SFPO is formulated as follows:

$$\mathcal{L}_{\text{SFPO}}(\pi_\theta; \pi_{\text{ref}}) = \mathcal{L}_{\text{F}}(\pi_\theta; \pi_{\text{ref}}) + \lambda \mathcal{L}_{\text{S}}(\pi_\theta; \pi_{\text{ref}}), \tag{5}$$

where π_θ denotes the model continuously optimized during post-training, while π_{ref} refers to OracleAnalyser-sft. The objective of $\mathcal{L}_{\text{F}}(\cdot;\cdot)$ aims to guide the model to reduce excessive attention to pairs that it struggles to rank. To achieve this, FocalPO (Liu et al., 2025b) is introduced as $\mathcal{L}_{\text{F}}(\cdot;\cdot)$, which is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{F}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\big[\\ p(y_w \succ y_l \mid x)^\gamma \log p(y_w \succ y_l \mid x)\big] + \text{c},\end{aligned} \tag{6}$$

where γ is a tunable focusing hyperparameter set to 0.05, and c denotes a constant term. When the implicit reward model correctly classifies a preference pair, the modulating factor is close to 1 and leaves the loss unchanged. Conversely, as $p(y_w \succ y_l \mid x) \to 0$, the factor approaches 0, thereby reducing its contribution to the objective. $\mathcal{L}_{\text{S}}(\cdot;\cdot)$, on the other hand, stabilizes training by constraining the divergence between the π_θ and π_{ref}. The formula is shown as follows:

$$\mathcal{L}_{\text{S}}(\pi_\theta; \pi_{\text{ref}}) = \mathbb{E}_{x\sim\mathcal{D}}\left[D_{\text{KL}}\left(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\right)\right], \tag{7}$$

here, we adopt the reverse KL divergence to prevent the policy model from deviating excessively from the reference model, thus ensuring stable training. Guided by SFPO, the model can focus on more valuable samples during the preference training stage and achieve stabler optimization, ultimately yielding OracleAnalyser.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	Acc
Qwen-VL-MAX	29.11	15.20	8.19	4.43	32.41	9.04	20.14	2.4
GLM-4.5V	28.11	12.32	6.57	3.65	31.68	6.53	20.02	3.2
SenseChat-Vision	27.91	12.48	6.44	3.70	31.87	6.66	19.68	3.8
Intern-S1	28.32	12.57	6.65	3.89	31.58	6.48	20.82	14.0
InternVL-3.5-241B	28.21	12.53	6.63	3.84	31.55	6.48	20.78	12.8
InternVL-3-78B	27.79	12.69	6.32	3.54	31.50	6.68	19.91	7.2
Qwen3-VL-30B	28.41	13.10	6.38	3.44	32.05	7.06	18.94	1.2
Qwen2.5-VL-72B	29.61	14.56	7.54	4.32	32.30	8.34	20.02	1.6
Qwen2.5-VL-7B	20.77	9.93	5.62	3.07	26.92	6.67	16.52	0.6
Qwen2.5-VL-3B	18.22	8.14	4.09	2.23	25.78	5.14	15.60	0.4
OracleAnalyser-sft	86.43	80.77	74.65	69.86	64.68	52.09	62.05	36.2
OracleAnalyser	**87.92**	**81.96**	**76.23**	**72.77**	**67.86**	**55.07**	**65.42**	**39.6**

Table 1: Performance comparison of different models on the in-domain test set to evaluate their capabilities in oracle bone analysis and recognition. OracleAnalyser is trained based on Qwen2.5-VL-3B. Acc denotes recognition accuracy, and the best results are highlighted in **bold**.

4 Experiments

4.1 Proposed Benchmark

Existing benchmarks are limited to the recognition task and cannot evaluate model's analysis capability for oracle bone scripts. However, such analysis ability is crucial and even plays a key role in deciphering these ancient symbols. To address this limitation, we propose a new benchmark designed to assess the reasoning capability of MLLMs in oracle bone analysis. Our benchmark includes both in-domain and out-of-domain evaluations.

The in-domain evaluation aims to assess the model's fundamental reasoning and recognition capabilities. This evaluation is conducted on an in-domain test set containing 500 samples. To examine the model's ability to analyze oracle bone images, we adopt BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L as evaluation metrics. Specifically, BLEU-1 and ROUGE-1 measure the accuracy of key terms; BLEU-2 and ROUGE-2 reflect local semantic consistency; BLEU-3 and BLEU-4 evaluate sentence-level coherence; while ROUGE-L assesses global semantic consistency. In addition to analysis performance, we also adopt recognition accuracy to evaluate the model's final recognition ability.

The out-of-domain evaluation is designed to test the model's generalization ability when encountering unseen oracle bone scripts. This evaluation uses an out-of-domain test set containing approximately 500 samples. For metrics, recognition accuracy is replaced with the Sentence-Bert-Score(SBS), whose formulation is given as follows:

$$\text{SBS} = \frac{1}{N}\sum_{i=1}^{N}\text{cosine}(p_i, l_i) \tag{8}$$

where N denotes the number of samples in the dataset, and cosine represents the cosine similarity. p_i refers to the representation of the MLLM's output encoded by Sentence-Bert (Wang et al., 2020), while l_i denotes the representation of the corresponding ground-truth encoded by Sentence-Bert. This metric is used to evaluate the similarity between the model's output and the ground truth, and a higher similarity indicates that the model's output are closer to the ground truth, thus contributing more effectively to oracle bone decipherment.

4.2 Implementation Details

In our experimental setup, we train the model using a single NVIDIA A100 GPU. During the SFT stage, OracleAnalyser-sft is fine-tuned on top of Qwen2.5-VL-3B with the LoRA strategy. The LoRA rank is set to 8, and all layers of the model are fine-tuned. The batch size is 32, and the model is trained for 100 epochs using the AdamW optimizer with a learning rate of 5e-5. In the SFPO stage, the model continues training based on OracleAnalyser-sft, with a batch size of 8 and 20 training epochs. The learning rate is set to 5e-6, and the balancing coefficient λ for $\mathcal{L}_{\text{F}}$ and $\mathcal{L}_{\text{S}}$ is 0.5. In $\mathcal{L}_{\text{F}}$, the parameters β and γ are set to 0.1 and 0.05, respectively.

4.3 In-domain and Out-of-domain Evaluation

We evaluate the performance of our proposed method and several the most competitive MLLMs

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	SBS
Qwen-VL-MAX	29.18	14.05	8.16	4.45	32.37	8.99	20.16	69.98
GLM-4.5V	28.09	12.33	6.53	3.67	31.65	6.52	19.93	68.92
SenseChat-Vision	27.88	12.48	6.35	3.69	31.89	6.67	19.65	69.02
Intern-S1	28.37	12.60	6.66	3.86	31.58	6.49	20.83	68.88
InternVL-3.5-241B	28.27	12.52	6.61	3.84	31.58	6.46	20.81	68.81
InternVL-3-78B	27.83	12.99	6.09	3.39	31.06	6.80	19.49	67.92
Qwen3-VL-30B	28.47	13.13	6.20	3.32	31.48	7.08	18.57	60.34
Qwen2.5-VL-72B	28.63	14.42	7.58	4.35	32.42	8.38	20.04	68.75
Qwen2.5-VL-7B	20.36	8.91	4.15	3.04	26.85	5.92	16.33	58.97
Qwen2.5-VL-3B	18.81	7.94	3.71	2.08	25.08	4.70	15.22	49.99
OracleAnalyser-sft	76.77	67.99	57.75	49.54	38.81	18.27	33.82	69.22
OracleAnalyser	**77.61**	**68.58**	**58.31**	**50.13**	**39.85**	**19.44**	**35.01**	**70.35**

Table 2: Performance comparison of different models on the out-of-domain test set to evaluate their analytical capability when facing the unseen oracle bone characters. The best results are highlighted in **bold**.

on both in-domain and out-of-domain test sets. The comparison models include Qwen-VL, InternVL, GLM, and SenseChat series. It is worth noting that, apart from Qwen2.5-VL-3B, which serves as our baseline, all comparison models have significantly larger parameter scales than OracleAnalyser. For reproducibility, we fix the random seed to 42.

For the in-domain evaluation, the results are shown in Table 1. It is evident that OracleAnalyser significantly outperforms other models across all metrics, even those with much larger parameter scales. Specifically, for oracle bone analysis, OracleAnalyser achieves improvements ranging from 35% to 70% across different metrics. And for oracle bone recognition, it yields gains exceeding 20%. Moreover, we observe that SFT alone already brings substantial performance improvements, indicating that a well-designed supervised fine-tuning stage is essential for effective oracle bone analysis and recognition. Building upon this foundation, SFPO further enhances performance through preference-based optimization, allowing the model to learn more discriminative and stable reasoning patterns. These results demonstrate that, despite having fewer parameters, OracleAnalyser achieves superior and more robust performance in both analysis and recognition of oracle bone scrits.

For the out-of-domain evaluation, the results are presented in Table 2. Similarly, OracleAnalyser achieves the best performance across all metrics. Compared with the in-domain test, its performance on oracle bone analysis decreases, which is expected due to unseen samples. Nevertheless, OracleAnalyser still maintains a 10%-50% improvement over other models. In terms of SBS, OracleAnalyser achieves the highest score, indicating that its generated analyses are semantically closer to the ground truth and thus more beneficial for oracle bone deciphering. These results demonstrate that when confronted with unseen oracle bone characters, OracleAnalyser exhibits even stronger analytical and interpretive abilities than other powerful MLLMs. This confirms that, through post-training, our model has acquired a generalized reasoning capability for oracle bone analysis.

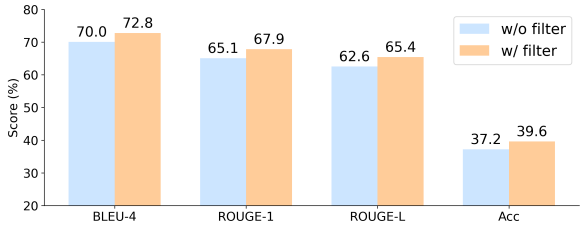

Figure 4: Ablation study on whether to filter the preference data.

4.4 Ablation Study

Ablation study on whether to filter the preference data. To evaluate the impact of filtering preference data on the final model performance, we conduct the following ablation study. The filtering process represents removing ambiguous preference pairs that are difficult to distinguish. Before filtering, the preference dataset contained over 4700 pairs, while after filtering, it was reduced to approximately 3400 pairs. OracleAnalyser-sft model is trained by SFPO using both datasets respectively, and the results are shown in Figure 4. It is evident that the OracleAnalyser trained with the filtered preference data achieves better performance across multiple metrics. Specifically, the model trained on

Training	BLEU-4	ROUGE-1	ROUGE-L	Acc
SFT	69.86	64.68	62.05	36.2
SFT+DPO	70.84	65.77	63.19	39.4
SFT+$\mathcal{L}_{\mathrm{F}}$	72.46	67.65	65.09	**39.8**
SFT+$\mathcal{L}_{\mathrm{F}}$+$\mathcal{L}_{\mathrm{S}}$	**72.77**	**67.86**	**65.42**	39.6

Table 3: Ablation study on different training strategies.

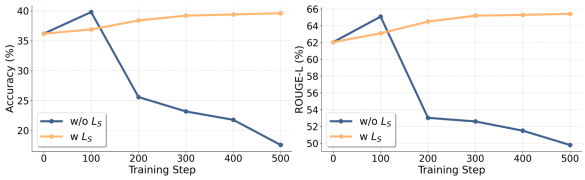

Figure 5: Ablation study on whether to employ $\mathcal{L}_{\mathrm{S}}$

the filtered data shows improvements of 2.8% in BLEU-4, ROUGE-1, and ROUGE-L, and a 2.4% gain in recognition accuracy. These results demonstrate that, in the field of oracle bone analysis, a larger amount of preference data does not necessarily lead to better performance. In fact, filtering out indistinguishable preference pairs can indeed enhance the effectiveness of preference optimization.

Ablation study on different training strategies. We conduct the ablation study to examine the impact of different training strategies on the final performance. The experiment is performed on the in-domain test set, and the results are presented in Table 3. It can be observed that applying DPO for subsequent preference training on top of SFT leads to some improvement, but the gain is smaller compared to using $\mathcal{L}_{\mathrm{F}}$. This indicates that the indistinguishable preference pairs in the oracle bone preference dataset have a negative influence on DPO training. In contrast, $\mathcal{L}_{\mathrm{F}}$ demonstrates greater robustness by reducing excessive attention to pairs that are hard to rank. Furthermore, the model jointly optimized with both $\mathcal{L}_{\mathrm{F}}$ and $\mathcal{L}_{\mathrm{S}}$ achieves the best overall performance, validating the effectiveness of the proposed SFPO framework.

Ablation study on whether to employ $\mathcal{L}_{\mathrm{S}}$. We conduct this ablation study to evaluate the effect of $\mathcal{L}_{\mathrm{S}}$ on model performance. The experiment is carried out on the in-domain test set, using recognition accuracy and ROUGE-L as evaluation metrics. As shown in Figure 5, the model with $\mathcal{L}_{\mathrm{S}}$ performs slightly worse at the early stage compared to the one trained with only $\mathcal{L}_{\mathrm{F}}$. However, the model trained solely with $\mathcal{L}_{\mathrm{F}}$ quickly reaches a peak and then suffers a sharp performance drop as training continues, eventually leading to collapse. In contrast, the model constrained by $\mathcal{L}_{\mathrm{S}}$ exhibits much

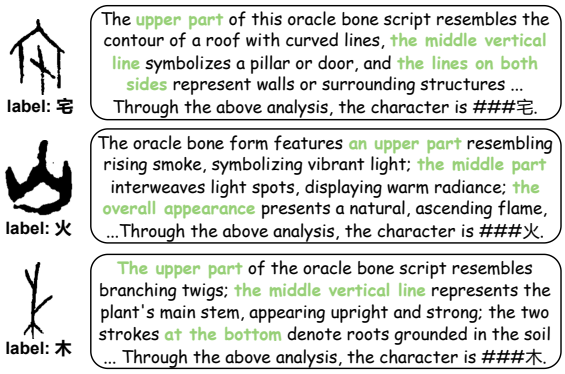

Figure 6: Visualization of OracleAnalyser outputs.

greater stability, and through the joint effect of $\mathcal{L}_{\mathrm{S}}$ and $\mathcal{L}_{\mathrm{F}}$, it achieves superior and more consistent performance in the later stages. Therefore, we conclude that $\mathcal{L}_{\mathrm{S}}$ effectively stabilizes training by preventing π_θ from deviating excessively from π_{ref}, thereby mitigating the impact of noisy data and enhancing the preference optimization process.

4.5 Visual Analysis

To better understand OracleAnalyser's reasoning behavior, we visualize its outputs in Figure 6. The results show that OracleAnalyser can accurately perceive and decompose oracle bone glyphs, and associate their components with real-world entities. This indicates that the model has learned to infer hidden semantics from pictographic structures. Evidently, through explicit reasoning and multi-stage post-training, OracleAnalyser demonstrates fundamental pictographic analysis capabilities which are valuable for oracle bone decipherment.

5 Conclusion

In this paper, we first identify the limitations of existing approaches and benchmarks in the oracle bone field, which focus mainly on recognition accuracy while neglecting equally crucial analytical reasoning. To address this, we propose OracleAnalyser, a post-training-based reasoning framework for oracle bone analysis. We also release an oracle bone reasoning dataset, an oracle bone preference dataset, and a new benchmark to evaluate the MLLM's analytical capabilities on oracle bone scripts. Extensive experiments demonstrate that OracleAnalyser, with only 3B parameters, achieves significantly superior analytical performance compared to much larger MLLMs. In the future, we will focus on out-of-domain oracle bone analysis, further contributing to the decipherment and understanding of oracle bone scripts.

6 Limitations

Although we have successfully developed OracleAnalyser, a post-training-based reasoning framework for oracle bone analysis, and released new oracle bone reasoning and preference datasets, as well as a newly designed benchmark for evaluating the analytical performance of MLLMs on oracle bone scripts, our work still has several limitations.

For OracleAnalyser, in out-of-domain evaluations, while the model can provide valuable analytical insights to assist in the decipherment of unseen oracle bone scripts, it is still unable to directly recognize characters that have never been encountered before. From a data perspective, the vast number and uneven quality of ancient oracle-related documents pose significant challenges. Although we carefully curated and integrated high-quality data, achieving comprehensive coverage of all oracle bone domains remains difficult. Our datasets are primarily collected from online sources and inevitably contain noise. Moreover, the current dataset scale is relatively limited; future work could focus on building a larger and more information-rich corpus. As for the benchmark, the evaluation metrics for oracle bone analysis can be further refined, and more diverse assessment perspectives may be explored to provide a more comprehensive understanding of model performance.

7 Ethical Considerations

Copyright and Licensing Statement. We strictly adhere to all copyright and licensing requirements associated with the datasets used in our work and ensure that their usage fully complies with the corresponding license agreements. When releasing related datasets and models, we commit to following all licensing terms to maintain legality.

Data Privacy. All datasets used in this study are publicly available and contain no personal or sensitive user information. Therefore, the OracleAnalyser model, along with the released reasoning dataset, preference dataset, and benchmark, poses no concerns regarding data privacy.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zhangquan Chen, Xufang Luo, and Dongsheng Li. 2025a. Visrl: Intention-driven visual perception via reinforced reasoning. *Preprint*, arXiv:2503.07523.

Zijian Chen, Tingzhu Chen, Wenjun Zhang, and Guangtao Zhai. 2024. Obi-bench: Can lmms aid in study of ancient script on oracle bones? *arXiv preprint arXiv:2412.01175*.

Zuyao Chen, Jinlin Wu, Zhen Lei, Marc Pollefeys, and Chang Wen Chen. 2025b. Compile scene graphs with reinforcement learning. *arXiv preprint arXiv:2504.13617*.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. *Preprint*, arXiv:2503.17352.

Haisu Guan, Jinpeng Wan, Pengjie Wang, Kaile Zhang, Zhebin Kuang, Xinyu Wang, Han Shengwei, Yongge Liu, Xiang Bai, Lianwen Jin, and Yuliang Liu. 2024a. An open dataset for the evolution of oracle bone characters: Evobc. *arXiv 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 model. In *Proceedings of the 62th Annual Meeting of the Association for Computational Linguistics*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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.

Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. 2025. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *Preprint*, arXiv:2503.16188.

Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, Tianwen Qian, Bin Zhu, Na Zhao, and Yu-Gang Jiang. 2024. Look before you decide: Prompting active deduction of mllms for assumptive reasoning. *arXiv preprint arXiv:2404.12966*.

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, and 1 others. 2025a.

A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*.

Tong Liu, Xiao Yu, Wenxuan Zhou, Jindong Gu, and Volker Tresp. 2025b. Focalpo: Enhancing preference optimizing by focusing on correct preference rankings. *arXiv preprint arXiv:2501.06645*.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. 2025c. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*.

Kaixin Peng, Mengyang Zhao, Haiyang Yu, Teng Fu, and Bin Li. 2025. Interpretable oracle bone script decipherment through radical and pictographic analysis with lvlms. *arXiv preprint arXiv:2508.10113*.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Jiapeng Wang, Yifan Zhang, Zhuoma GongQue, Chong Sun, Yida Xu, Yadong Xue, and 1 others. 2025. V-oracle: Making progressive reasoning in deciphering oracle bones for you and me. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20124–20150.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *Preprint*, arXiv:2504.07615.

Haoyuan Sun, Jiaqi Wu, Bo Xia, Yifu Luo, Yifei Zhao, Kai Qin, Xufei Lv, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. 2025. Reinforcement fine-tuning powers reasoning capability of multimodal large language models. *Preprint*, arXiv:2505.18536.

Kimi Team, Angang Du, Flood Sung, Zhilin Yang, and 1 others. 2025a. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

Kimi Team, Angang Du, Flood Sung, Zhilin Yang, and 1 others. 2025b. Kimi-vl technical report. *Preprint*, arXiv:2504.07491.

Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, and 1 others. 2025. A survey on post-training of large language models. *arXiv e-prints*, pages arXiv–2503.

Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. 2025a. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*.

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

Peiyu Wang, Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, and 1 others. 2025b. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*.

Pengjie Wang, Kaile Zhang, Yuliang Liu, Jinpeng Wan, Haisu Guan, Zhebin Kuang, Xinyu Wang, Lianwen Jin, and Xiang Bai. 2024a. An open dataset for oracle bone script recognition and decipherment. *arXiv preprint arXiv:2401.15365*.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, and 1 others. 2025c. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Zhicong Wu, Qifeng Su, Ke Gu, and Xiaodong Shi. 2024. A cross-font image retrieval network for recognizing undeciphered oracle bone inscriptions. *arXiv preprint arXiv:2409.06381*.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*.

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*.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. 2025. Perception-r1: Pioneering perception policy with reinforcement learning. *Preprint*, arXiv:2504.07954.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.