

# V-Oracle: Making Progressive Reasoning in Deciphering Oracle Bones for You and Me

Runqi Qiao<sup>1\*†</sup>, Qiuna Tan<sup>1\*†</sup>, Guanting Dong<sup>1</sup>, Minhui Wu<sup>2</sup>, Jiapeng Wang<sup>3‡</sup>,  
Yifan Zhang<sup>1</sup>, Zhuoma GongQue<sup>1</sup>, Chong Sun<sup>2</sup>, Yida Xu<sup>1</sup>, Yadong Xue<sup>1</sup>,  
Ye Tian<sup>1</sup>, Zhimin Bao<sup>2</sup>, Lan Yang<sup>1‡</sup>, Chen Li<sup>2</sup>, Honggang Zhang<sup>1</sup>

<sup>1</sup>Beijing University of Posts and Telecommunications    <sup>2</sup>WeChat Vision, Tencent Inc.

<sup>3</sup>South China University of Technology

{qrq, qiunatan, dongguanting, ylan, zhhg}@bupt.edu.cn

chaselli@tencent.com

## Abstract

Oracle Bone Script (OBS) is a vital treasure of human civilization, rich in insights from ancient societies. However, the evolution of written language over millennia complicates its decipherment. In this paper, we propose **V-Oracle**, an innovative framework that utilizes Large Multi-modal Models (LMMs) for interpreting OBS. V-Oracle applies principles of pictographic character formation and frames the task as a visual question-answering (VQA) problem, establishing a multi-step reasoning chain. It proposes a multi-dimensional data augmentation for synthesizing high-quality OBS samples, and also implements a multi-phase oracle alignment tuning to improve LMMs' visual reasoning capabilities. Moreover, to bridge the evaluation gap in the OBS field, we further introduce **Oracle-Bench**, a comprehensive benchmark that emphasizes process-oriented assessment and incorporates both standard and out-of-distribution setups for realistic evaluation. Extensive experimental results can demonstrate the effectiveness of our method in providing quantitative analyses and superior deciphering capability.

## 1 Introduction

Ancient scripts witness the evolution of human civilization. Among them, China's oldest pictographic script, Oracle Bone Script (OBS), reveals the culture and life of the Shang Dynasty (1600 to 1046 BCE) (Keightley, 1979). These scripts record complex social structures and cover aspects such as agriculture and trade (Boltz, 1986), providing important clues for tracing the roots of feudal dynasties along the Yellow River. In recent years, deciphering OBS has been a pursuit for historians. However, due to the passage of time and the inherent linguistic complexities, only about 2,200 of

\*Equal contribution

†Work done as intern at WeChat, Tencent Inc.

‡Corresponding author

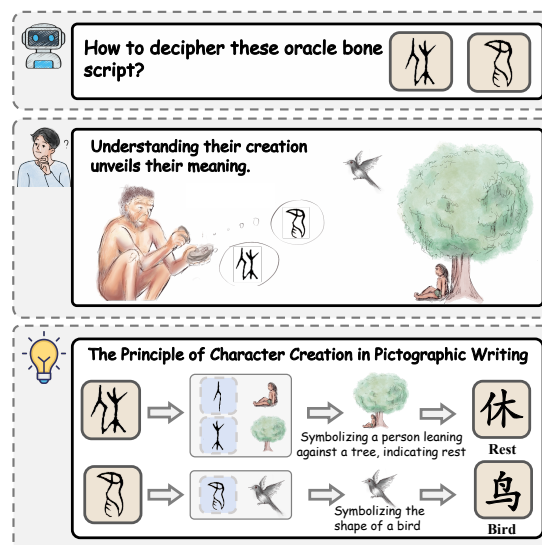


Figure 1: Exploring how the ancients transformed real-world visuals into oracle bone script can help us better decipher its meaning.

the approximately 4,500 known characters have been successfully deciphered, presenting an ongoing challenge for both historians and scientists.

Deciphering oracle bone scripts using AI technology is not an overnight task. Most studies focus primarily on OBS detection (Liu et al., 2020; Huang et al., 2024; Weng et al., 2024) and recognition (Han et al., 2020a; Wang and Deng, 2024; Wang et al., 2023) fields. With the rapid development of diffusion models, several efforts offer possibilities for deeper OBS exploration. Recently, some methods attempt to model the entire OBS deciphering process in an end-to-end manner (Chang et al., 2022; Guan et al., 2024a). These efforts lay a solid foundation for the digitalization and recognition of OBS, though they still face the following challenges:

**(1) Significant Visual Differences:** As written language has evolved, Chinese characters have significantly changed in shape, making it suboptimal to rely solely on modern character visual informa-

tion for deciphering.

**(2) Neglecting the Principles of Pictographic Character Formation:** As shown in Figure 1, the ancients create OBS by recording elements of real scenes. Directly allowing models to reverse-predict oracle bone images from modern character images, while ignoring the unique pictographic features of oracle bone script, is inconsistent with the principles of character formation.

In this paper, we aim to start from the principles of ancient character formation and validate that LMMs should understand the principles of oracle bone script to achieve better deciphering capabilities. Specifically, we emphasize that LMMs should first decompose and understand the pictographic meanings of various OBS radicals, and then progressively infer the corresponding modern Chinese characters. Based on this, we propose **V-Oracle**, a progressive framework incorporating OBS deciphering logic chain. Departing from conventional approaches, our method reformulates the deciphering task into a visual question-answering (VQA) paradigm (Antol et al., 2015; Karpathy and Fei-Fei, 2017). We collect OBS samples and the corresponding explanations from authoritative Chinese character databases, systematically breaking down each inscription into a complete deciphering logic chain. Building upon this foundation, our framework automatically generates over 50 different task types following directed acyclic graph rules, subsequently applying a multi-dimensional data augmentation strategy to simulate real-world deciphering scenarios. To further enhance the deciphering capability of LMMs, we also propose a multi-phase oracle alignment tuning which consists of LLM Pre-Alignment, Visual Alignment, and Cross-Modal Alignment. Through this progressive training process, our model can understand OBS from single/multi-modal and coarse/fine-grained information.

Moreover, to bridge the gap in automated evaluation of the OBS field, we introduce **Oracle-Bench**, a VQA benchmark focused on the comprehensive assessment of LMMs’ OBS deciphering abilities. The benchmark structures evaluation tasks into two forms: interpretation and deciphering of OBS, which include both standard and out-of-distribution settings. This process-oriented design enables systematic assessment of LMMs’ generalized abilities for real-world OBS decipherment scenarios.

In summary, our contributions are as follows:

- We propose V-Oracle, a progressive framework for LMMs’ OBS decipherment through two core innovations: a novel data synthesis framework with a multi-dimensional augmentation strategy, and a multi-stage oracle alignment tuning, progressing from the basic understanding of OBS to radical decomposition and visual enhancement.
- We introduce Oracle-Bench, a VQA benchmark dedicated to comprehensively assessing LMMs’ OBS decipherment. Comprising 2,834 samples across 13 subfields and 3 character formation principles, it evaluates process-oriented reasoning while covering both standard and OOD challenges.
- Extensive experimental results demonstrate the superiority of our V-Oracle. We also hope that the public availability of our V-Oracle and Oracle-Bench along with the quantitative analysis can significantly contribute to the deciphering of OBS.

## 2 Related Work

Deciphering oracle bone scripts is a complex process that has attracted considerable interest from researchers employing modern AI technologies to explore these ancient texts. Most studies concentrate on OBS detection and recognition using images of inscriptions (Liu et al., 2020; Huang et al., 2024; Xing et al., 2019; Weng et al., 2024; Meng et al., 2018; Han et al., 2020a; Wang and Deng, 2024; Wang et al., 2023; wang et al., 2020; Ge et al., 2021; Gao and Liang, 2020). With advancements in diffusion models, various methods provide fresh insights and opportunities for a deeper exploration of oracle bone decipherment (Chang et al., 2022; Guan et al., 2024a). Meanwhile, OBS-focused benchmark (Chen et al., 2024b) evaluates the current abilities of LMMs in whole-process OBS tasks, while exploring feasible solutions for future OBS research. Recently, several methods have emerged to model the entire OBS deciphering process in an end-to-end manner, such as employing a conditional diffusion strategy to convert oracle bone scripts directly into modern Chinese character (Guan et al., 2024b), interpreting ancient texts by decomposing and reconstructing characters from different historical periods (Wang et al., 2024c). However, these approaches often overlook the critical role of pictographic semantics in reconstructing the logic of OBS decipherment. More

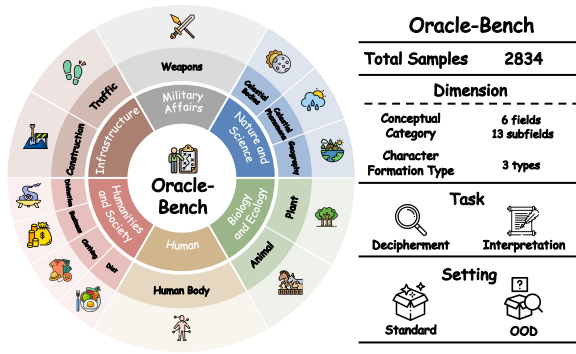


Figure 2: The overview diagram and categories of Oracle Bone Script in Oracle-Bench.

detailed introductions of related works are given in supplementary materials.

### 3 Oracle-Bench

#### 3.1 Task Definition

Due to the absence of a VQA task paradigm for oracle bone script, our primary goal is to establish such a task while ensuring scholarly rigor. Given oracle bone script  $X$ , the task is to decipher it into modern Chinese characters  $Y$ . This involves navigating the complexities of ancient pictographic symbols and accurately translating them into their modern equivalents.

#### 3.2 Design Principles

Oracle-Bench provides a comprehensive evaluation framework by adhering to the following principles: **Multi-dimensional coverage.** It incorporates key domains and character formation principles to ensure a comprehensive evaluation of OBS.

**Professional pictographic explanations.** Each oracle bone script is accompanied by pictographic interpretations annotated by human experts.

**Process-oriented assessment.** Leveraging the pictographic meanings of oracle bone script, we provide process supervision to evaluate its interpretation mechanism.

**Alignment with real deciphering scenarios.** It includes standard and challenging out-of-distribution scenarios, ensuring that the difficulty of deciphering oracle bone script matches real-world contexts.

#### 3.3 Data Collection

To construct a high-quality and rigorous Oracle-Bench, as shown in Figure 2, we mainly follow the following steps:

**Basic Collection:** As oracle bone script is a rigorous field of study, the images and their cor-

Table 1: The statistics of our Oracle-Bench.

Statistic	Number
Total Questions	2,834
<b>OBS Interpretation</b>	
Total questions	1,417
Proportion of answer A	364 (25.6%)
Proportion of answer B	351 (24.8%)
Proportion of answer C	347 (24.5%)
Proportion of answer D	355 (25.1%)
<b>OBS Decipherment</b>	
Total questions	1,417
Proportion of answer A	378 (26.7%)
Proportion of answer B	342 (24.1%)
Proportion of answer C	352 (24.8%)
Proportion of answer D	345 (24.3%)
<b>Average Question Length</b>	
OBS Interpretation	221
OBS Decipherment	138
<b>Max Question Length</b>	
OBS Interpretation	286
OBS Decipherment	138

responding Chinese characters in Oracle-Bench are primarily sourced from public datasets (Wang et al., 2024d; Li et al., 2024a), with additional data crawled and processed from authoritative websites<sup>1</sup>. The dataset draws upon authenticated research institutions to ensure the reliability of our data.

**Pictographic Annotation:** Based on the professional OBS websites and books referenced in (Wang et al., 2024d), we retrieve and collect the corresponding pictographic interpretations of modern Chinese characters on these platforms. Additionally, we conduct cross-validation across three websites to ensure the consistency and quantity of these interpretations with OBS.

**Domain Classification:** To facilitate research across various fields, such as civic organizations, museums, and academic institutions, we categorize OBS by key domains and character formation principles, allowing researchers to easily locate relevant sections while ensuring a structured representation of OBS. As shown in Figure 2, we group these characters and define categories accordingly.

Note that, we strictly follow the licenses of the mentioned data, all of which are released under appropriate Creative Commons (CC) licenses. And we strictly comply with copyright and licensing rules, ensuring that we refrain from using data from sites that forbid copying and redistribution.

<sup>1</sup><https://humanum.arts.cuhk.edu.hk/Lexis/lexi-mf/>

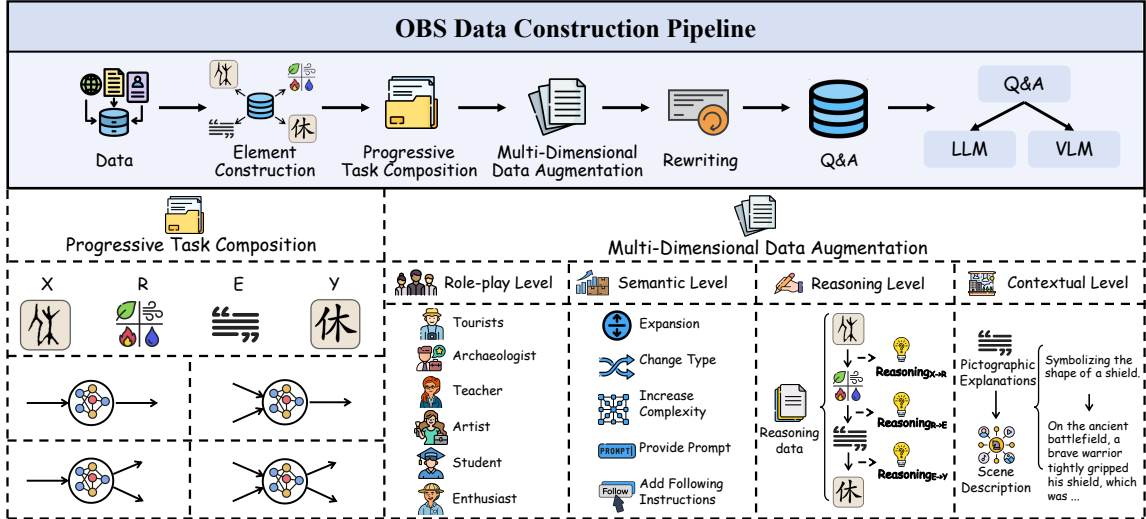


Figure 3: The overall framework of OBS Data Construction Pipeline, Progressive Task Composition, and Multi-Dimensional Data Augmentation in our V-Oracle.

### 3.4 Data Statistics and Evaluation Protocol

Oracle-Bench is a comprehensive VQA benchmark for evaluating LMM’s OBS deciphering ability, encompassing 13 subfields across 6 domains and 3 character formation principles, totaling 2,834 samples. Each sample includes expert-level pictographic explanations. The evaluation tasks are divided into two parts: **OBS Interpretation** indicates that given oracle bone script  $X$ , the goal is to identify their pictographic meanings  $E$ ; and **OBS Decipherment** aims to identify the corresponding modern Chinese character  $Y$  given oracle bone script  $X$ . The statistics of Oracle-Bench are shown in Table 1.

For each question, we format it as a multiple-choice question (with one correct option and three random options). Additionally, for the evaluation scenarios, we establish both standard and out-of-distribution (OOD) setups to fit real-world oracle bone script decipherment scenarios. **Standard** means that the evaluation includes cases where the deciphered modern Chinese characters are the same as those inferred from the train set, but the OBS images are entirely different; and **Out-Of-Distribution (OOD)** indicates that the evaluation of OBS images, corresponding pictographic explanations, and inferred modern Chinese characters has no overlap with the training data. Deciphering any out-of-distribution characters in this challenging setup is highly valuable.

Thus, through the combined metrics, we could evaluate whether the LMMs truly possesses the generalized ability for deciphering OBS.

## 4 V-Oracle

### 4.1 Oracle Synthetic Data Construction

**Training Data Organization.** The foundational OBS data for V-Oracle primarily comes from web-crawled sources. We construct the data sources using the GPT-4o (OpenAI, 2024), indexing it by modern Chinese characters.

**Progressive Task Composition.** We model the four elements — oracle bone text  $X$ , pictographic radicals  $R$ , pictographic meanings  $E$ , and modern Chinese characters  $Y$  — as a node set  $V$  and construct  $N$  directed acyclic graphs (DAGs). As shown in Figure 3, any node in the set  $V$  can be designated as the head node  $u$ , with other specific nodes as the tail node  $v$ , forming a directed acyclic graph  $(u, v)$ . Based on this concept, we define  $D(n)$  different subgraphs as  $D(n) = \sum_{k=0}^n \binom{n}{k} \cdot D(k-1) \cdot D(n-k)$ . Considering the in-degree and out-degree characteristics of the DAG, we can classify oracle bone VQA input-output pairs into four scenarios: single input & single output, single input & multiple outputs (e.g., input  $X$ , outputs  $(R, E, Y)$ ), multiple inputs & single output, and multiple inputs & multiple outputs. With varying task compositions, our downstream LMMs can better capture the relationships among the four key elements in the deciphering logic chain represented in the DAG, enhancing their understanding of the mapping between different nodes during reasoning.

**Multi-Dimensional Data Augmentation.** Given the limited data in the OBS field, adding various



supervisory signals is vital for enhancing LMM decryption capabilities. Thus, we have designed a multi-dimensional data augmentation strategy (as shown in Figure 3) to strengthen data quality at the following levels, with an example of the reasoning level prompt template shown in Figure 4 and the complete set of templates provided in the supplementary:

**1) Role-playing Level.** The design intention of V-Oracle is to assist various groups involved in oracle bone script research, such as archaeologists and history teachers, in deciphering the script. To cater to these different groups, we apply role-playing enhancements to the task-combined oracle bone data for distinct roles. As shown in Figure 3, we input each sample along with role information into the supervision model, which then generates relevant data based on the provided identity.

**2) Semantic Level.** We treat the deciphering of OBS as a multi-modal reasoning task and introduce 5 semantic augmentation methods as listed in Figure 3. Given a question, we inject augmentation requirements as contextual information and use the robust supervision model GPT-4o to generate outputs with a temperature of 1. We adopt greedy decoding, ensuring that each augmentation generates one corresponding augmented question.

**3) Reasoning Level.** To alleviate the gap in the fine-grained logical connections during the decoding process, we instruct GPT-4o to provide additional chain-of-thought explanations for each reasoning link under the given conditions  $(X, R, E, Y)$  to enhance the logical connections.

**4) Contextual Level.** We prompt GPT-4o to use a template to associate and expand on these scenes based on the pictographic meanings of oracle bone images. Considering the oracle bone decoding chain, we propose that the supervision model enhances scenes from both global and radical perspectives: global perspective allows GPT-4o to intuitively rewrite pictographic meanings, boldly imagining and describing a complete scene while staying true to the character; radical perspective enables GPT-4o to interpret the radical’s pictographic elements based on the meanings and the described scene. For instance, the character "宀" can be understood as "house". Our contextualization enables us to reverse-engineer the observational scenes recorded by the ancients according to their character creation principles, enriching the contextual information of the pictographic meanings.

<Template> Reasoning Level

Now you are an oracle bone script expert. I will provide you with a set of oracle bone script data that contains logical connections between the components. Your task is to attempt to deduce the logical relationships among these elements.

Specifically, this set of data includes an oracle bone script image, the corresponding modern Chinese character, the radicals into which the oracle bone script can be decomposed, and the pictographic meaning of the script.

You need to provide the following key reasoning steps:

Reason1: Why can this oracle bone script image be decomposed into these radicals?  
Reason2: Why can this oracle bone script and its radicals correspond to this pictographic meaning?  
Reason3: Why can this oracle bone script, in combination with the radicals and pictographic meaning, be deduced to form the corresponding modern Chinese character?

Please adhere to the following response format:

<Reasoning Template>

Figure 4: The template for the prompt in the Reasoning Level section of Mutli-Dementional data augmentation.

## 4.2 Oracle Alignment Tuning

After augmentation, we obtain the synthetic dataset  $D$ . To further differentiate the training data for the visual encoder and LLM backbone, we use GPT-4o to classify data formats, determining if the input requires information from the OBS image  $x$  for accurate reasoning. If it does, it is classified as multimodal data; if not, as text data. Ultimately, we derive two datasets:  $D_t$  and  $D_m$ .

Here we detail the progressive multi-stage training process of V-Oracle (as shown in Figure 5):

**Stage1: LLM Pre-Alignment.** Current LLMs have demonstrated scalable performance in reasoning with synthetic augmented data. Considering the scarcity of high-quality multimodal oracle-related data, we first pre-align the LLM backbone in the LMM. With the synthesized dataset  $D_t$ , we perform supervised fine-tuning on the LLM backbone:

$$\mathcal{L}(\theta) = \sum_{x \in D_t} \log P_{\theta}(y_n | \text{prompt}(x)), \quad (1)$$

where  $\log P(\cdot)$  denotes the probability distribution of the LLM’s output,  $\theta$  represents the model parameters, and  $x$  indicates the inputs of  $D_t$ .

**Stage2: Visual Alignment.** To assist the visual encoder capture fine-grained information in OBS, we design a two-stage coarse-to-fine visual alignment strategy. For coarse-grained visual alignment, we follow Equation (1), replacing  $D_t$  with  $D_m$  to conduct visual supervised fine-tuning on the encoder, establishing an initial visual perception

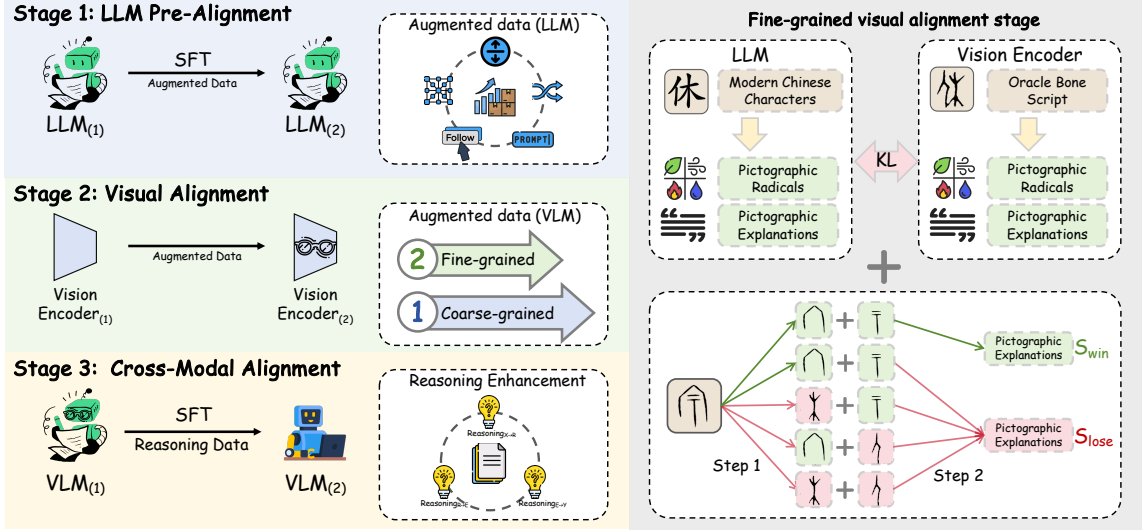


Figure 5: Overview of our Oracle Alignment Tuning. Left: The multi-stage training process of V-Oracle. Right: Detailed illustration of fine-grained visual alignment.

of OBS. Then, we introduce a fine-grained visual alignment stage by proposing an **oracle step-wise Direct Preference Optimization (DPO)** to differentiate similar radicals and glyphs, while adhering to the prior conditions of the step-by-step derivation of oracle bone meanings ( $X \rightarrow R \rightarrow E$ ):

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left( \beta \left( \log \frac{\pi(y^+ | x; s_{\text{win}})}{\pi(y^- | x; s_{\text{lose}})} - \log \frac{\pi_{\text{ref}}(y^+ | x; s_{\text{win}})}{\pi_{\text{ref}}(y^- | x; s_{\text{lose}})} \right) \right), \quad (2)$$

$$s_{\text{lose}} = \begin{cases} ((X \rightarrow R)_{\text{lose}}, (R \rightarrow E)_{\text{win}}), \\ ((X \rightarrow R)_{\text{win}}, (R \rightarrow E)_{\text{lose}}), \\ ((X \rightarrow R)_{\text{lose}}, (R \rightarrow E)_{\text{lose}}), \end{cases} \quad (3)$$

where the set  $s_{\text{lose}}$  captures instances of sampling errors at any step. We further stabilize this process by using the pre-aligned LLM from the first stage as a teacher model. We align their logits using KL divergence, as shown in Equation (4), ensuring stability in the reasoning process:

$$\mathcal{L}_{\text{kl}} = D_{\text{KL}}(l \parallel v) = \sum_i l_i \log \frac{l_i}{v_i}, \quad (4)$$

where  $v$  and  $l$  represent the logits of the LLM with frozen LLM params and the pre-aligned LLM. Therefore, the total loss is computed as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{DPO}} + (1 - \alpha) \cdot \mathcal{L}_{\text{kl}}. \quad (5)$$

Notably, we freeze the parameters of the LLM and only fine-tune the visual encoder throughout the entire Stage 2<sup>2</sup>.

<sup>2</sup>More details can be found in supplementary materials.

**Stage3: Cross-Modal Alignment.** In the final phase, we jointly train the visual encoder with the LLM backbone for overall alignment. Unlike the previous stage, we retain only the multimodal data from dataset  $D_m$ , where the task format includes OBS input  $X$  and outputs  $\{R, E, Y\}$ . This is to ensure that: (1) the LLM treats oracle bone script deciphering as a general multimodal reasoning task, gradually deriving answers from the deciphering chain; (2) to prevent interference from other task formats in a multi-task setup. Thus, our multimodal instruction tuning effectively transforms the oracle bone script deciphering task into a reasoning task format, introducing additional reasoning steps and information while optimizing the use of both textual and visual data, enhancing performance in OBS decipherment.

## 5 Experiment

**Evaluation Models and Metric.** We examine the performance of foundation models across two distinct categories on Oracle-Bench: (a) Closed-source LLMs: GPT-4o (OpenAI, 2024), GPT-4V (OpenAI, 2023); (b) Open-source LLMs: InternVL2.5-78B, InternVL2.5-8B, InternVL2.5-4B, InternVL2.5-1B (Chen et al., 2024a), Qwen2.5-VL-72B, Qwen2.5-VL-7B (Team, 2025), GLM-4V-9B (GLM et al., 2024), LongVA (Zhang et al., 2024), MiniCPM-V 2.6 (Yao et al., 2024), DeepSeek-VL-7B (Lu et al., 2024), Phi-3.5-vision-instruct (Abdin et al., 2024). For automated evaluation, we standardize all samples into a

Table 2: The performance of different LMMs on Oracle-Bench. The best results are highlighted in **bold**, **blue** for the best closed-source model, and **green** for the best open-source model. (Acc: Accuracy, HS: Humanities and Society, Hum: Human, MA: Military Affairs, Inf: Infrastructure, BE: Biology and Ecology, NS: Nature and Science).

Models	Decipherment						Interpretation							
	Acc (↑)	HS (↑)	Hum (↑)	MA (↑)	Inf (↑)	BE (↑)	NS (↑)	Acc (↑)	HS (↑)	Hum (↑)	MA (↑)	Inf (↑)	BE (↑)	NS (↑)
<i>Closed-source</i>														
GPT-4o	41.00%	48.11%	39.49%	34.57%	42.60%	41.73%	40.59%	44.32%	47.03%	44.57%	48.77%	48.52%	37.97%	42.57%
GPT-4V	29.43%	35.68%	26.33%	27.78%	31.36%	32.33%	26.24%	39.52%	47.57%	36.72%	41.36%	39.65%	42.11%	33.17%
<i>Open-source</i>														
InternVL2.5-78B	43.47%	43.24%	39.03%	46.30%	43.79%	47.74%	45.05%	47.71%	50.81%	45.73%	51.23%	48.52%	50.00%	42.57%
Qwen2.5-VL-72B	36.91%	38.92%	33.49%	38.89%	45.56%	37.59%	32.67%	42.27%	47.57%	40.42%	51.23%	43.20%	37.22%	40.10%
GLM-4V-9B	31.76%	30.81%	30.02%	30.86%	37.87%	34.59%	28.22%	32.25%	38.92%	38.34%	25.93%	34.32%	25.94%	24.75%
InternVL2.5-8B	35.99%	35.68%	36.95%	33.95%	33.14%	35.71%	38.61%	41.78%	45.95%	42.73%	43.83%	46.15%	36.84%	37.13%
LongVA	28.09%	30.27%	31.41%	27.16%	28.99%	22.93%	25.74%	32.11%	31.89%	29.10%	30.86%	37.87%	35.71%	30.20%
MiniCPM-V 2.6	28.02%	33.51%	26.79%	27.78%	24.85%	25.19%	32.18%	35.92%	42.16%	34.87%	35.19%	33.73%	38.72%	31.19%
Qwen2.5-VL-7B	30.28%	34.59%	24.25%	30.86%	34.32%	34.21%	30.20%	36.49%	37.84%	34.64%	45.06%	34.91%	38.72%	30.69%
DeepSeek-VL-7B	23.71%	21.08%	23.09%	27.78%	29.59%	22.18%	21.29%	26.11%	25.95%	29.33%	25.93%	30.18%	25.19%	17.33%
Phi-3.5-vision-instruct	21.24%	22.70%	18.94%	18.52%	25.44%	25.19%	18.32%	24.14%	21.62%	25.17%	25.93%	22.49%	24.44%	23.76%
InternVL2.5-4B	32.46%	32.43%	31.18%	40.12%	44.38%	22.93%	31.68%	36.34%	44.32%	38.11%	38.89%	36.69%	27.07%	35.15%
Qwen2.5-VL-3B	24.28%	24.32%	22.17%	32.72%	26.04%	23.68%	21.29%	32.46%	34.05%	29.79%	40.12%	36.69%	33.08%	26.24%
InternVL2.5-1B	28.44%	24.32%	30.95%	32.10%	33.14%	24.06%	25.74%	30.28%	35.68%	26.79%	39.51%	34.32%	27.82%	25.25%
<b>V-Oracle (Qwen2-7B)</b>	<b>81.65%</b>	<b>74.05%</b>	<b>79.21%</b>	<b>85.19%</b>	<b>85.90%</b>	<b>80.45%</b>	<b>89.11%</b>	<b>87.58%</b>	<b>82.16%</b>	<b>84.53%</b>	<b>92.59%</b>	<b>91.72%</b>	<b>87.97%</b>	<b>91.09%</b>

multiple-choice format. We use regex to match the LMMs’ predictions and then calculate their accuracy against the ground-truth answers for main results.

**Training Details of V-Oracle.** We adopt our multi-stage progressive training method with the following experimental setup:

- **LLM Pre-Alignment.** We performed supervised fine-tuning on the Qwen2 (Wang et al., 2024b) language model using 216K Oracle bone script corpus, with a learning rate of  $2 \times 10^{-6}$ , training for 3 epochs.
- **Fine-grained Visual Alignment.** We selected SigLIP (Zhai et al., 2023) as the pre-trained visual encoder (resolution 384x384) and used a two-layer MLP as the projection layer. Building on this, we used the pre-aligned LLM from the first stage to sequentially perform visual alignment. The two steps during this period used 87k and 80k visual problems respectively, each training for one epoch. Moreover, the default value of the hyperparameter  $\alpha$  is 0.9 and  $\beta$  is 0.1.
- **Cross-Modal Alignment.** We performed full-parameter fine-tuning for 1 epoch using 30k OBS QA pairs with reasoning paths. All experiments were conducted on 8 NVIDIA

A800-SXM4-80GB GPUs. The optimizer used was AdamW with a warm-up ratio of 0.03 and a cosine learning rate decay strategy. The batch sizes for the three stages were 32, 8, and 8, respectively.

## 5.1 Main Results.

**V-Oracle significantly outperforms advanced competitors.** Table 2 reports the evaluation results on Oracle-Bench, covering both closed-source and open-source MLLMs. V-Oracle (Qwen2-7B) establishes a new state-of-the-art (SOTA) among all MLLMs. Notably, despite having only 7B parameters, it surpasses models with up to 78B parameters, highlighting its remarkable efficiency in OBS decipherment and interpretation.

**Advanced LMMs still face challenges in the OBS domain.** All existing LMMs perform poorly in both OBS interpretation and decipherment tasks. Specifically, powerful closed-source models like GPT-4V and GPT-4o consistently achieve accuracy rates below 50% on both tasks; the best open-source model, InternVL2.5-78B, also shows similar results, highlighting the challenging nature of tasks in Oracle-Bench. In the specialized domain, GPT-4o has the lowest accuracy in the military domain (only 34%), possibly due to significant differences in pictographic meanings and character forms between ancient and modern culture.

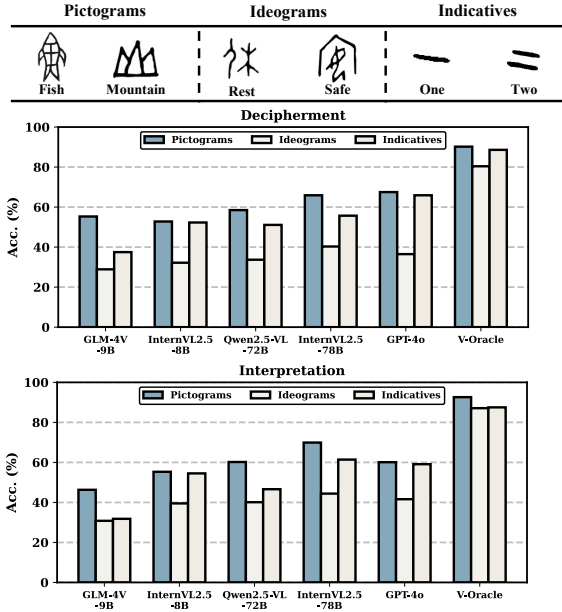


Figure 6: Performance of LMMs across character formation principles.

Table 3: Ablation & OOD study for V-Oracle.

Models	Standard		OOD	
	Deciph.	Interp.	Deciph.	Interp.
GPT-4o	41.00%	44.32%	40.70%	41.71%
InternVL2.5-78B	42.48%	44.18%	40.95%	43.72%
<b>V-Oracle (Qwen2)</b>	<b>81.65%</b>	<b>87.58%</b>	<b>59.80%</b>	<b>69.35%</b>
w/o Visual Alignment (Fine)	79.75%	84.12%	59.55%	64.32%
w/o Visual Alignment (Coarse & Fine)	66.90%	75.02%	43.22%	51.26%
w/o Cross Modal Alignment	51.45%	75.86%	39.70%	63.32%
w/o Multi-Dimensional Data Augmentation	44.04%	64.43%	32.66%	49.75%

**The performance of LMMs on OBS interpretation generally outperforms the decipherment.** This finding aligns with our intuition. The reasoning path for the interpretation task ( $X \rightarrow R$ ) is shorter than that for the decipherment task  $\{X \rightarrow (R, E, Y)\}$ , reflecting the difference in problem-solving difficulty. Moreover, the candidate answers in the interpretation task provide explanations for the oracle bone script, offering LMMs more supervisory information.

**LMMs excel at pictograms but struggle with ideograms.** As shown in Figure 6, most LMMs perform best on pictograms, as their direct visual cues are easier to interpret. Indicatives yield moderate accuracy due to their simpler structures, while ideograms pose the greatest challenge, requiring the integration of multiple pictographic elements to convey abstract meanings. This highlights the need to enhance pictographic understanding, which could improve generalization across all character types in OBS decipherment.

**Larger parameter sizes lead to more stable improvement in OBS field.** Focusing on the laws of parameter scaling in InternVL2.5 families, we can observe that larger parameter scales in LLMs generally yield better performance, indicating that parameter size in the text decoder is crucial for achieving generalization in OBS reasoning.

## 5.2 Quantitative Analysis

**Ablation Study.** To investigate the roles of different stages in V-Oracle, we conduct an ablation study, with results shown in Table 3. Versions without specific modules are denoted as *w/o*. Key findings include: 1) Each alignment stage significantly impacts performance, confirming the necessity of all design components. 2) Cross-modal alignment and multi-dimensional data augmentation are essential, as removing either significantly degrades performance in OBS decipherment and interpretation. This highlights the importance of integrating multi-modal reasoning and data augmentation to enhance model robustness and generalization. 3) Removing coarse-grained alignment causes a greater performance drop than fine-grained alignment, while their combination yields the best results. This aligns with our expectations, as coarse-grained visual fine-tuning establishes a foundation by capturing global visual information, while fine-grained DPO and KL divergence enhance the model’s ability to distinguish challenging samples with similar radicals. These findings support our coarse-to-fine design rationale.

**OOD Analysis.** Table 3 also presents the results from the challenging OOD setting. Compared to the in-domain evaluations, the OOD setting indeed poses significant challenges for V-Oracle, reinforcing the notion that the two metrics in Oracle-Bench exhibit a clear difficulty gradient. Notably, fine-tuning the V-Oracle framework solely with Qwen2-7B significantly surpasses existing strong models, both open-source and closed-source, particularly showing a 28% improvement over GPT-4o in OBS interpretation. We have also found that V-Oracle demonstrates the ability to decipher characters through pictographic understanding, while open-source models rely solely on direct character mapping. This finding highlights that V-Oracle improves the generalization of LMMs in real-world OBS decipherment. More qualitative examples are provided in supplementary materials.

**Process-oriented Analysis.** To analyze the relationship between OBS interpretation and decipher-



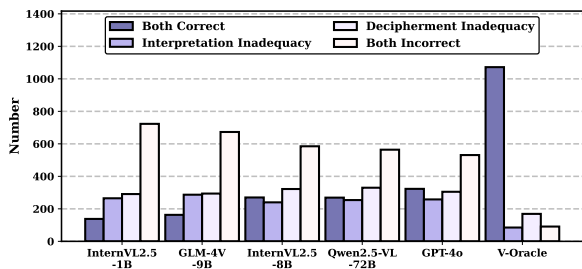


Figure 7: Process-oriented analysis of different LMMs. For a given sample, based on OBS deciphering and interpretation results, we can classify them into different conditions.

ment, we construct a four-dimensional metric in Figure 7. In cases of “Both Correct” and “Both Incorrect”, V-Oracle (Qwen2-7B) achieve the highest consistency of correct answers (1072) and the fewest errors (91), demonstrating the strong generalization across all parameter sizes of LMMs. Further deeper analysis can reveal that the number of correct answers in "Interpretation inadequacy" is generally lower than in "Decipherment", suggesting that deciphering OBS is more challenging for LMMs than interpretation, reinforcing the rationale behind our oracle bone decipherment chain  $\{X \rightarrow (R, E, Y)\}$  and emphasizing that LMMs should first grasp pictographic meanings before reasoning to modern characters. These results indicate that most errors in existing LMMs arise from mistakes in the deciphering process rather than a lack of capability.

## 6 Conclusion

In this paper, we reformulate OBS decipherment as a VQA paradigm and establish a deciphering logic chain. We introduce V-Oracle, a progressive framework that enhances OBS decipherment. Through task composition and multi-dimensional data augmentation, our three-stage alignment paradigm enables LMMs to progress from basic understanding to advanced reasoning. To address evaluation gaps, we also propose Oracle-Bench, the first VQA benchmark for rigorously assessing LMMs in OBS decipherment. Extensive experiments can demonstrate V-Oracle’s strong generalization. We hope that V-Oracle and Oracle-Bench can promote the development of future research in the OBS field.

## Acknowledgements

We thank the reviewers for their valuable comments. This work was partially supported by

STI2030-Major Projects (2021ZD0200600).

## Limitations

Despite our best efforts to model the oracle bone decipherment process by faithfully replicating the ancient methods of character creation and integrating the indispensable elements of radicals and pictographic meanings, the establishment of **V-Oracle** and **Oracle-Bench** has several limitations:

**V-Oracle.** Firstly, the vast amount of ancient literature related to oracle bones and its varied quality present a significant challenge. While we have carefully curated and synthesized high-quality data, it remains impossible to cover all relevant domains of oracle bones. Most of our dataset comes from internet sources and includes primarily deciphered oracle bone images, but some texts remain difficult to collect and verify. During the training process, we incorporated strategies to maximize the use of available data and mitigate the impact of missing information. Notably, our method has demonstrated effectiveness in out-of-domain (OOD) scenarios. We hope V-Oracle can introduce new ideas and offer fresh perspectives for the oracle bone research community.

**Oracle-Bench.** In constructing Oracle-Bench, we prioritized the use of trusted and authoritative data sources to ensure its reliability. However, this also led to the exclusion of certain data due to issues like multiple interpretations for the same character. Additionally, while efforts were made to standardize the dataset format and include high-quality images, minor randomness might still exist. In the future, we plan to periodically review and correct these excluded data points and continue refining Oracle-Bench to expand its data coverage and improve its overall accuracy.

## Ethical Considerations

**Copyright and Licensing.** We strictly comply with the copyright requirements of all datasets used and ensure their usage aligns with the respective licensing agreements. When publishing related data and models, we guarantee adherence to all dataset licensing terms to maintain legality and compliance.

**Data Privacy.** As all datasets we use are open-source and do not involve any personal user information, **V-Oracle** and **Oracle-Bench** do not pose any concerns regarding data privacy.

## References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. **VQA: visual question answering**. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- William G. Boltz. 1986. **Early chinese writing**. *World Archaeology*, 17:420–436.
- Xiang Chang, Fei Chao, Changjing Shang, and Qiang Shen. 2022. **Sundial-gan: A cascade generative adversarial networks framework for deciphering oracle bone inscriptions**. In *Proceedings of the 30th ACM International Conference on Multimedia, MM '22*, page 1195–1203, New York, NY, USA. Association for Computing Machinery.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.
- Zijian Chen, Tingzhu Chen, Wenjun Zhang, and Guangtao Zhai. 2024b. Obi-bench: Can llms aid in study of ancient script on oracle bones? *arXiv preprint arXiv:2412.01175*.
- Guanting Dong, Xiaoshuai Song, Yutao Zhu, Runqi Qiao, Zhicheng Dou, and Ji-Rong Wen. 2024. Toward general instruction-following alignment for retrieval-augmented generation. *arXiv preprint arXiv:2410.09584*.
- Junheng Gao and Xun Liang. 2020. **Distinguishing oracle variants based on the isomorphism and symmetry invariances of oracle-bone inscriptions**. *IEEE Access*, 8:152258–152275.
- Wenyang Ge, Guoying Liu, and Jing Lv. 2021. **Oracle bone inscriptions extraction by using weakly supervised instance segmentation under deep network**. In *2021 IEEE 4th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pages 229–233.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucien Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuntao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. **Chatglm: A family of large language models from glm-130b to glm-4 all tools**. *Preprint, arXiv:2406.12793*.
- 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.
- Haisu Guan, Huanxin Yang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. 2024c. **Deciphering oracle bone language with diffusion models**. *arXiv preprint arXiv:2406.00684*.
- Wenhui Han, Xinlin Ren, Hangyu Lin, Yanwei Fu, and Xiangyang Xue. 2020a. **Self-supervised learning of orc-bert augmentor for recognizing few-shot oracle characters**. In *Computer Vision – ACCV 2020: 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 – December 4, 2020, Revised Selected Papers, Part VI*, page 652–668, Berlin, Heidelberg. Springer-Verlag.
- Xu Han, Yuzhuo Bai, Keyue Qiu, Zhiyuan Liu, and Maosong Sun. 2020b. **Isobs: An information system for oracle bone script**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 227–233.
- Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. 2024. **mplug-docowl 1.5: Unified structure learning for ocr-free document understanding**. *arXiv preprint arXiv:2403.12895*.

- Ting Huang, Junya Liu, Xin Zhou, and Zhen Yang. 2024. [Oracle recognition based on improved yolov8](#). In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering, CAICE '24*, page 671–675, New York, NY, USA. Association for Computing Machinery.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025a. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.
- Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. 2025b. [Flashrag: A modular toolkit for efficient retrieval-augmented generation research](#). In *Companion Proceedings of the ACM on Web Conference 2025, WWW '25*, page 737–740, New York, NY, USA. Association for Computing Machinery.
- Andrej Karpathy and Li Fei-Fei. 2017. [Deep visual-semantic alignments for generating image descriptions](#). *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676.
- David N. Keightley. 1979. [The shang state as seen in the oracle-bone inscriptions](#). *Early China*, 5:25–34.
- Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2023. Instructer: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint arXiv:2309.11911*.
- Bang Li, Donghao Luo, Yujie Liang, Jing Yang, Zeng-mao Ding, Xu Peng, Boyuan Jiang, Shengwei Han, Dan Sui, Peichao Qin, et al. 2024a. Oracle bone inscriptions multi-modal dataset. *arXiv preprint arXiv:2407.03900*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024b. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024c. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.
- Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.
- Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yongkang Wu, Zhonghua Li, Qi Ye, and Zhicheng Dou. 2024d. Retrollm: Empowering large language models to retrieve fine-grained evidence within generation. *arXiv preprint arXiv:2412.11919*.
- Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024e. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26763–26773.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Guoying Liu, Jici Xing, and Jing Xiong. 2020. [Spatial pyramid block for oracle bone inscription detection](#). In *Proceedings of the 2020 9th International Conference on Software and Computer Applications, ICSCA '20*, page 133–140, New York, NY, USA. Association for Computing Machinery.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024d. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. [Deepseek-vl: Towards real-world vision-language understanding](#). *Preprint*, arXiv:2403.05525.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.
- Lin Meng, Naoki Kamitoku, and Katsuhiko Yamazaki. 2018. [Recognition of oracle bone inscriptions using deep learning based on data augmentation](#). In *2018 Metrology for Archaeology and Cultural Heritage (MetroArchaeo)*, pages 33–38.
- OpenAI. 2024. [Hello gpt-4o](#).
- R OpenAI. 2023. Gpt-4v (ision) system card. *Citekey: gptvision*.
- Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. 2024a. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.




- Runqi Qiao, Lan Yang, Kaiyue Pang, and Honggang Zhang. 2024b. 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.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.
- Xiaoshuai Song, Muxi Diao, Guanting Dong, Zhengyang Wang, Yujia Fu, Runqi Qiao, Zhexu Wang, Dayuan Fu, Huangxuan Wu, Bin Liang, et al. 2024. Cs-bench: A comprehensive benchmark for large language models towards computer science mastery. *arXiv preprint arXiv:2406.08587*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Qwen Team. 2025. [Qwen2.5-vl](#).
- Mei Wang and Weihong Deng. 2024. A dataset of oracle characters for benchmarking machine learning algorithms. *Scientific Data*, 11(1):87.
- Mei Wang, Weihong Deng, Jiani Hu, and Sen Su. 2024a. Unsupervised attention regularization based domain adaptation for oracle character recognition. *arXiv preprint arXiv:2409.15893*.
- Mei Wang, Weihong Deng, and Cheng-Lin Liu. 2022. Unsupervised structure-texture separation network for oracle character recognition. *IEEE Transactions on Image Processing*, 31:3137–3150.
- Mei Wang, Weihong Deng, and Sen Su. 2023. [Oracle character recognition using unsupervised discriminative consistency network](#). *Preprint*, arXiv:2312.06075.
- Nan wang, Zhenquan Zhao, and Qingju Jiao. 2020. [Oracle bone inscriptions components analysis based on image similarity](#). In *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, volume 9, pages 1666–1670.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. 2024c. [Puzzle pieces picker: Deciphering ancient chinese characters with radical reconstruction](#). *Preprint*, arXiv:2406.03019.
- Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Jinpeng Wan, Haisu Guan, Zhebin Kuang, Lianwen Jin, Xiang Bai, et al. 2024d. An open dataset for oracle bone script recognition and decipherment. *arXiv preprint arXiv:2401.15365*.
- Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.
- Xinying Weng, Yifan Li, Shuaidong Hao, and Jialiang Hou. 2024. [Oracle bone script similiar character screening approach based on simsiam contrastive learning and supervised learning](#). *Preprint*, arXiv:2408.06811.
- Jici Xing, Guoying Liu, and Jing Xiong. 2019. [Oracle bone inscription detection: a survey of oracle bone inscription detection based on deep learning algorithm](#). In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing, AIIPCC ’19*, New York, NY, USA. Association for Computing Machinery.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.
- Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. 2024. Internlm-math: Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332*.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.
- Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.
- Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024. [Long context transfer from language to vision](#). *Preprint*, arXiv:2406.16852.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Sid-dhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.
- Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. 2024. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. *arXiv preprint arXiv:2408.08640*.



## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Related Work</b>	<b>2</b>
<b>3</b>	<b>Oracle-Bench</b>	<b>3</b>
3.1	Task Definition . . . . .	3
3.2	Design Principles . . . . .	3
3.3	Data Collection . . . . .	3
3.4	Data Statistics and Evaluation Protocol . . . . .	4
<b>4</b>	<b>V-Oracle</b>	<b>4</b>
4.1	Oracle Synthetic Data Construction	4
4.2	Oracle Alignment Tuning . . . . .	5
<b>5</b>	<b>Experiment</b>	<b>6</b>
5.1	Main Results. . . . .	7
5.2	Quantitative Analysis . . . . .	8
<b>6</b>	<b>Conclusion</b>	<b>9</b>
<b>A</b>	<b>More Details on V-Oracle</b>	<b>13</b>
A.1	Models Architecture . . . . .	13
A.2	Details of Multi-Dimensional Data Augmentation . . . . .	13
<b>B</b>	<b>More Details on Oracle-Bench</b>	<b>13</b>
B.1	Details of the Evaluated Models . . . . .	13
B.2	Details of the Model Hyperparameters . . . . .	14
<b>C</b>	<b>Additional Results of V-Oracle</b>	<b>14</b>
C.1	Case Study . . . . .	14
C.2	Detailed Results of Ablation Studies	14
<b>D</b>	<b>Additional Results on Oracle-Bench</b>	<b>16</b>
D.1	Details of Model Performance . . . . .	16
D.2	Detailed Performance on Each Category . . . . .	16
<b>E</b>	<b>Comparison with Existing Work</b>	<b>16</b>
E.1	Differences from OBSD . . . . .	16
E.2	Differences from Other Studies . . . . .	18
E.3	Comparative Analysis of Oracle-Bench and OBI-Bench . . . . .	18
<b>F</b>	<b>Detailed Related Work</b>	<b>19</b>
<b>G</b>	<b>Broaden Impact</b>	<b>19</b>

 <Template> Semantic Level

Now you are an expert in rewriting oracle bone script-related questions. I will provide you with a set of questions and answers about oracle bone script, and your task is to rewrite each set of questions and answers according to the following requirements:

- **Rewrite 1- Expand**  
Try to expand the questions and answers. My original questions may be brief as they follow a unified template, so you need to enhance the understandability by expanding them.
- **Rewrite 2- Change the question type**  
Try to change the question type. You can attempt different formats such as Q&A, multiple-choice, or fill-in-the-blank. If it's a multiple-choice question, please supplement it with complete options.
- **Rewrite 3- Increase complexity**  
Increase the complexity of the question. Depending on the nature of the question, you can try reducing the provided conditions or making the question more challenging.
- **Rewrite 4- Provide more prompts**  
Based on the additional data I provide, give more hints for the question. Additional data: {Example}.
- **Rewrite 5- Increase the difficulty of following instructions**  
Try to impose some instruction-following requirements on the original question, such as limiting the answer length or specifying a particular format for the response. Please ensure that the answer follows the given instruction.

Finally, please follow the format below:  
<Template >

Figure 8: The template for the prompt in the Semantic Level section of Mutli-Dementional data augmentation.

## A More Details on V-Oracle

### A.1 Models Architecture

V-Oracle is initialized with carefully selected components to ensure strong multimodal capabilities. Specifically, we adopt Qwen2-7B (Wang et al., 2024b) as the LLM due to its advanced natural language understanding and generation capabilities. For the visual encoder, we utilize a SigLIP-so400m-patch14-384 architecture, following the LLaVA (Liu et al., 2024b) framework to achieve effective vision-language alignment.

### A.2 Details of Multi-Dimensional Data Augmentation

As shown in Figure 8 to 10, we present the prompt templates used in the Multi-Dimensional Data Augmentation (*Role-play level, Semantic level, Contextual level, Reasoning level*).

## B More Details on Oracle-Bench

### B.1 Details of the Evaluated Models

To evaluate the performance of various multimodal large models in oracle bone decipherment and character interpretation, we selected the latest versions. Table 7 presents the release dates and specific sources of these models. Furthermore, we have

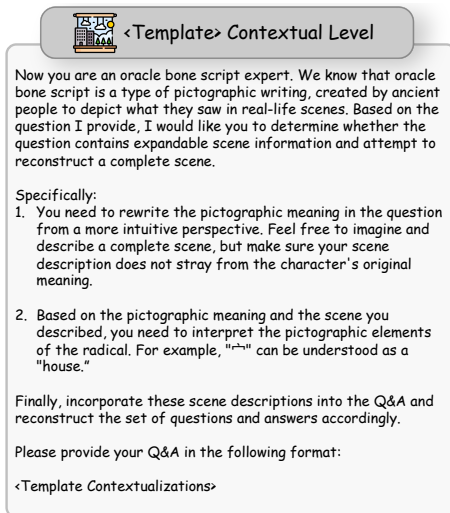


Figure 9: The template for the prompt in the Contextual Level section of Mutli-Dementional data augmentation.

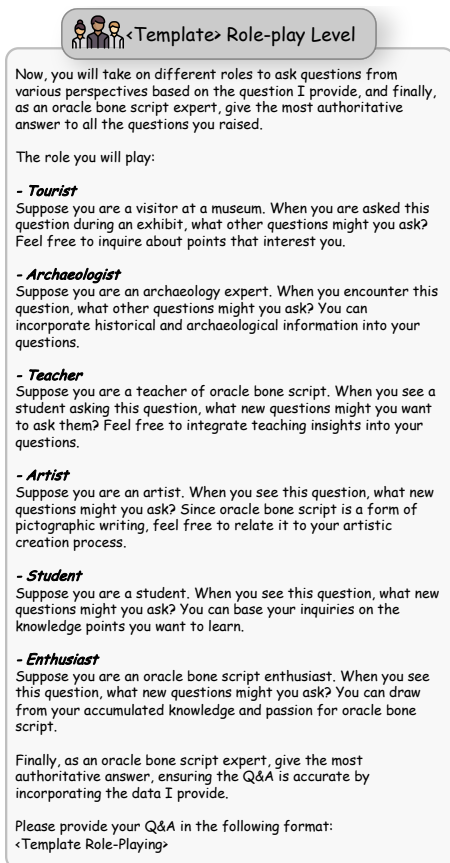


Figure 10: The template for the prompt in the Role-play Level section of Mutli-Dementional data augmentation.

showcased their architectural designs in Table 6 to provide a more comprehensive evaluation.

## B.2 Details of the Model Hyperparameters

For all closed-source models with API access, we use a standard generation approach and run the inference on CPUs, which typically completes within

a day. For open-source models, we utilize a cluster with 8 NVIDIA A800-SXM4-80GB GPUs to perform the inference, following the hyper-parameter settings specified in the model's inference samples. If no specific instructions are provided, we use the default settings. The specific generation parameters are detailed in Table 4 and 5.

## C Additional Results of V-Oracle

### C.1 Case Study

**Analysis in Standard Setting.** We present the comprehensive evaluation and response process in Figure 11 and Figure 12. The upper part of the figure illustrates the detailed evaluation workflow and prompt design of Oracle-Bench, while the lower part shows the response results of V-Oracle. In the example shown in Figure 11, V-Oracle accurately describes the pictographic meaning of the oracle bone script for "Light" and deduces its corresponding modern Chinese character based on the correct principles of character formation.

In contrast, Figure 12 presents a failure case using the "Mountain" example, where V-Oracle incorrectly identifies it as "Hill". To analyze the underlying reason for this error, we include an oracle bone script sample for "Hill" in the figure. Our observation reveals a high visual similarity between the two scripts, which makes it challenging for the model to differentiate such cases. As the first framework focusing on the decipherment of oracle bone scripts based on character formation principles, V-Oracle establishes a solid foundation. At the same time, by presenting these failure cases, we also provide new perspectives for future research to explore improved strategies for visual recognition.

### C.2 Detailed Results of Ablation Studies

As shown in Table 8 and Table 9, we present more detailed ablation study results. It can be observed that, whether in the OOD setup or the Standard setup, each layer of V-Oracle's design shows significant improvements in the results, further highlighting the importance of each layer.

**Analysis in OOD Setting.** Figure 13 illustrates the comparison between V-Oracle and open-source models in the OOD setting. Despite encountering unseen characters, V-Oracle attempts to infer correctness through pictographic understanding, while open-source models rely solely on direct character-to-modern-Chinese mapping, lacking an

Table 4: Generating parameters for Open-Source LMMs.

Model	Generation Setup
InternVL2.5-78B	do_sample = False, max_new_tokens = 1024
InternVL2-26B	do_sample = False, temperature = 0, max_new_tokens = 1024
InternVL2-8B	do_sample = False, temperature = 0, max_new_tokens = 1024
InternVL2-2B	do_sample = False, temperature = 0, max_new_tokens = 1024
InternVL2.5-8B	do_sample = False, max_new_tokens = 1024
InternVL2.5-4B	do_sample = False, max_new_tokens = 1024
InternVL2.5-1B	do_sample = False, max_new_tokens = 1024
Qwen2.5-VL-72B	do_sample = False, max_new_tokens = 1024
Qwen2-VL-72B	do_sample = False, temperature = 0, max_new_tokens = 1024
Qwen2.5-VL-7B	do_sample = False, max_new_tokens = 1024
Qwen2-VL-7B	do_sample = False, temperature = 0, max_new_tokens = 1024
Qwen2.5-VL-3B	do_sample = False, max_new_tokens = 1024
GLM-4V-9B	do_sample = False
Phi-3.5-vision-instruct	num_beams = 1, do_sample = False, max_new_tokens = 1024
MiniCPM-V 2.6	do_sample = True, max_length = 1024, top_k = 1
LongVA	do_sample = True, max_length = 1024, top_k = 1
DeepSeek-VL-7B	do_sample = True, max_length = 1024, top_k = 1

Table 5: Generating parameters for Closed-Source LMMs.

Model	Generation Setup
GPT-4o	"model" : "gpt-4o", "temperature" : 0, "max_tokens" : 1024
GPT-4V	"model" : "gpt-4-turbo", "temperature" : 0, "max_tokens" : 1024

Table 6: Model architecture of 13 LMMs evaluated on Oracle-Bench.

Models	LLM	Vision Encoder
GPT-4o	-	-
GPT-4V	-	-
InternVL2.5-78B	Qwen2.5-72B-Instruct	InternViT-6B-448px-V2_5
InternVL2-Llama3-76B	Hermes-2-Theta-Llama-3-70B	InternViT-6B-448px-V1-5
InternVL2-26B	InternLM2-chat-20b	InternViT-6B-448px-V1-5
InternVL2-8B	InternLM2_5-7b-chat	InternViT-300M-448px
InternVL2-2B	InternLM2-chat-1_8b	InternViT-300M-448px
InternVL2.5-8B	internlm2_5-7b-chat	InternViT-6B-448px-V2_5
InternVL2.5-4B	Qwen2.5-3B-Instruct	InternViT-300M-448px-V2_5
InternVL2.5-1B	Qwen2.5-0.5B-Instruct	InternViT-300M-448px-V2_5
Qwen2.5-VL-72B	Qwen2.5-72B-Instruct	CLIP ViT-bigG-P14
Qwen2.5-VL-7B	Qwen2.5-7B-Instruct	CLIP ViT-bigG-P14
Qwen2-VL-72B	Qwen2-72B	CLIP ViT-bigG-P14
Qwen2-VL-7B	Qwen2-7B	CLIP ViT-bigG-P14
GLM-4V-9B	GLM-9B	EVA_02_CLIP-E-P14 (4.7B)
Phi-3.5-vision-instruct	Phi-3-mini-128K-instruct	CLIP-ViT-L-P14-336
MiniCPM-V 2.6	Qwen2-7B	SigLLp-so400m-P14-980
LongVA	Qwen2-7B-Instruct	CLIP-ViT-L-P14-336
DeepSeek-VL-7B	DeepSeek-LLM-7B-base	SigLLp-L-P16-384 & SAM-B

understanding of the underlying formation principles. Notably, in this OOD scenario, the model has never been trained on any image data related to

this character. The fact that V-Oracle successfully deciphered it by leveraging pictographic meaning is a remarkable finding, further demonstrating that

Table 7: The release time and model source of LMMs used in Oracle-Bench

Model	Release Time	Source
GPT-4o (OpenAI, 2024)	2024-05	<a href="https://gpt4o.ai/">https://gpt4o.ai/</a>
GPT-4V (OpenAI, 2023)	2024-04	<a href="https://openai.com/index/gpt-4v-system-card/">https://openai.com/index/gpt-4v-system-card/</a>
InternVL2.5-78B (Chen et al., 2024a)	2024-12	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-78B">https://huggingface.co/OpenGVLab/InternVL2_5-78B</a>
InternVL2-Llama3-76B (Chen et al., 2024a)	2024-07	<a href="https://huggingface.co/OpenGVLab/InternVL2-Llama3-76B">https://huggingface.co/OpenGVLab/InternVL2-Llama3-76B</a>
InternVL2.5-8B (Chen et al., 2024a)	2024-12	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-8B">https://huggingface.co/OpenGVLab/InternVL2_5-8B</a>
InternVL2.5-4B (Chen et al., 2024a)	2024-12	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-4B">https://huggingface.co/OpenGVLab/InternVL2_5-4B</a>
InternVL2.5-1B (Chen et al., 2024a)	2024-12	<a href="https://huggingface.co/OpenGVLab/InternVL2_5-1B">https://huggingface.co/OpenGVLab/InternVL2_5-1B</a>
InternVL2-26B (Chen et al., 2024a)	2024-07	<a href="https://huggingface.co/OpenGVLab/InternVL2-26B">https://huggingface.co/OpenGVLab/InternVL2-26B</a>
InternVL2-8B (Chen et al., 2024a)	2024-07	<a href="https://huggingface.co/OpenGVLab/InternVL2-8B">https://huggingface.co/OpenGVLab/InternVL2-8B</a>
InternVL2-2B (Chen et al., 2024a)	2024-07	<a href="https://huggingface.co/OpenGVLab/InternVL2-2B">https://huggingface.co/OpenGVLab/InternVL2-2B</a>
Qwen2.5-VL-72B (Team, 2025)	2025-01	<a href="https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct">https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct</a>
Qwen2-VL-72B (Wang et al., 2024b)	2024-09	<a href="https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct">https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct</a>
Qwen2.5-VL-7B (Team, 2025)	2025-01	<a href="https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct">https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct</a>
Qwen2-VL-7B (Wang et al., 2024b)	2024-09	<a href="https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct">https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct</a>
Qwen2.5-VL-3B (Team, 2025)	2025-01	<a href="https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct">https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct</a>
GLM-4V-9B (GLM et al., 2024)	2024-06	<a href="https://huggingface.co/THUDM/glm-4v-9b">https://huggingface.co/THUDM/glm-4v-9b</a>
Phi-3.5-vision-instruct (Abdin et al., 2024)	2024-08	<a href="https://huggingface.co/microsoft/Phi-3.5-vision-instruct">https://huggingface.co/microsoft/Phi-3.5-vision-instruct</a>
MiniCPM-V 2.6 (Yao et al., 2024)	2024-08	<a href="https://huggingface.co/openbmb/MiniCPM-V-2_6">https://huggingface.co/openbmb/MiniCPM-V-2_6</a>
LongVA (Zhang et al., 2024)	2024-06	<a href="https://huggingface.co/lmms-lab/LongVA-7B">https://huggingface.co/lmms-lab/LongVA-7B</a>
DeepSeek-VL-7B (Lu et al., 2024)	2024-03	<a href="https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat/">https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat/</a>

understanding character formation principles can enhance a model’s generalization ability.

## D Additional Results on Oracle-Bench

### D.1 Details of Model Performance

Table 10 and Table 11 present additional experimental results of models on Oracle-Bench. Table 10 illustrates the model performance across six categories, while Table 11 presents the results based on the dimension of character construction. All of these results are in alignment with the conclusions drawn in the main text.

### D.2 Detailed Performance on Each Category

From Figure 15 to Figure 34, we further illustrate the detailed results under the second-level nodes in Oracle-Bench, providing a more granular analysis of model performance. Specifically, under the second-level nodes, the number of cases where interpretation outperforms decipherment exceeds those where decipherment outperforms interpretation across all models. This trend is particularly pronounced in V-Oracle. Moreover, by leveraging a stronger understanding of character formation principles, V-Oracle achieves significant improvements in decipherment compared to other models.

## E Comparison with Existing Work

### E.1 Differences from OBSD

OBSD (Guan et al., 2024c) is the first model to attempt deciphering oracle bone script, marking a

significant step forward in this challenging field. To better compare its approach with our work, **V-Oracle**, we aim to reproduce the results reported in the OBSD paper. However, OBSD does not release its model weights or provide detailed information about the allocation of its test set, making direct reproduction of its results infeasible. As a result, we compare the results indirectly while directly contrasting the methodologies and underlying motivations of each approach. This limitation is also why we do not present a direct comparison of OBSD’s results in the main text.

### Motivation & Method

- **Principles of Oracle Bone Script Formation vs. Evolutionary Mapping.** OBSD relies on the concept of Chinese character evolution, attempting to map the implicit correlation between oracle bone script and modern Chinese character distributions. However, the gap between these two character systems is significant, which may limit its performance. **V-Oracle**, on the other hand, emphasizes the principles of character formation, focusing on how ancient people originally created oracle bone script. By modeling the pictographic nature of these characters, **V-Oracle** trains the model to learn the compositional structure of each character, enabling more accurate and interpretable decipherment.
- **Integrated Model Design vs. Modular Com-**



Table 8: Detailed Ablation Study for V-Oracle under Standard Setup.

Model	Decipherment							Interpretation						
	Acc (↑)	HS (↑)	Hum (↑)	MA (↑)	Inf (↑)	BE (↑)	NS (↑)	Acc (↑)	HS (↑)	Hum (↑)	MA (↑)	Inf (↑)	BE (↑)	NS (↑)
<b>V-Oracle (Qwen2)</b>	81.65%	74.05%	79.21%	85.19%	85.80%	80.45%	89.11%	87.58%	82.16%	84.53%	92.59%	91.72%	87.97%	91.09%
w/o Visual Alignment (Fine)	79.75%	74.59%	77.14%	85.19%	84.02%	78.20%	84.16%	84.12%	76.76%	81.06%	89.51%	83.43%	87.97%	88.61%
w/o Visual Alignment (Coarse & Fine)	66.90%	60.54%	63.05%	67.28%	71.01%	70.68%	72.28%	75.02%	70.81%	69.75%	81.48%	76.33%	76.32%	82.18%
w/o Cross Model Alignment	51.45%	47.03%	49.65%	62.35%	54.44%	45.11%	56.44%	75.86%	68.11%	75.98%	78.40%	75.15%	77.07%	79.70%
w/o Visual & Cross Model Alignment	34.79%	29.19%	37.64%	36.42%	37.87%	31.20%	34.65%	42.20%	35.14%	46.19%	42.59%	39.05%	39.47%	46.04%
w/o Multi-Dimensional Data Augmentation	44.04%	40.54%	47.34%	42.59%	47.93%	43.23%	39.11%	64.43%	59.46%	64.20%	62.96%	63.91%	63.16%	72.78%

Table 9: Detailed Ablation Study for V-Oracle under OOD Setup.

Model	Decipherment							Interpretation						
	Acc (↑)	HS (↑)	Hum (↑)	MA (↑)	Inf (↑)	BE (↑)	NS (↑)	Acc (↑)	HS (↑)	Hum (↑)	MA (↑)	Inf (↑)	BE (↑)	NS (↑)
<b>V-Oracle (Qwen2)</b>	59.80%	51.52%	58.33%	61.54%	57.14%	55.95%	85.71%	69.35%	66.67%	63.64%	84.62%	68.57%	71.43%	73.81%
w/o Visual Alignment (Fine)	59.55%	59.09%	56.82%	69.23%	51.43%	58.33%	69.05%	64.32%	51.52%	56.82%	76.92%	57.14%	73.81%	83.33%
w/o Visual Alignment (Coarse & Fine)	43.22%	39.39%	41.67%	35.90%	42.86%	53.57%	40.48%	51.26%	43.94%	43.18%	64.10%	54.29%	54.76%	66.67%
w/o Cross Model Alignment	39.70%	33.33%	41.67%	58.97%	34.29%	34.52%	40.48%	63.32%	54.55%	68.18%	76.92%	42.86%	65.48%	61.90%
w/o Visual & Cross Model Alignment	24.62%	16.67%	30.30%	23.08%	8.57%	26.19%	30.95%	31.66%	27.27%	35.61%	33.33%	25.71%	26.19%	40.48%
w/o Multi-Dimensional Data Augmentation	32.66%	27.27%	39.39%	33.33%	31.43%	27.38%	30.95%	49.75%	42.42%	50.76%	53.85%	51.43%	45.24%	61.90%

Table 10: The performance of different LMMs on Oracle-Bench. The best results are highlighted in **bold**, **blue** for the best closed-source model, and **green** for the best open-source model. Metrics include Avg: Score<sub>average</sub>, Humanities and Society (HS), Human (Hum), Military Affairs (MA), Infrastructure (Inf), Biology and Ecology (BE), and Nature and Science (NS).

Model	Decipherment							Interpretation						
	Acc (↑)	HS (↑)	Hum (↑)	MA (↑)	Inf (↑)	BE (↑)	NS (↑)	Acc (↑)	HS (↑)	Hum (↑)	MA (↑)	Inf (↑)	BE (↑)	NS (↑)
<i>Closed-source</i>														
GPT-4o	41.00%	48.11%	39.49%	34.57%	42.60%	41.73%	40.59%	44.32%	47.03%	44.57%	48.77%	48.52%	37.97%	42.57%
GPT-4V	29.43%	35.68%	26.33%	27.78%	31.36%	32.33%	26.24%	39.52%	47.57%	36.72%	41.36%	39.65%	42.11%	33.17%
<i>Open-source</i>														
InternVL2.5-78B	43.47%	43.24%	39.03%	46.30%	43.79%	47.74%	45.05%	47.71%	50.81%	45.73%	51.23%	48.52%	50.00%	42.57%
InternVL2-26B	41.78%	42.16%	40.65%	48.77%	43.79%	42.48%	35.64%	40.08%	44.86%	42.03%	34.57%	48.52%	36.84%	33.17%
InternVL2.5-8B	35.99%	35.68%	36.95%	33.95%	33.14%	35.71%	38.61%	41.78%	45.95%	42.73%	43.83%	46.15%	36.84%	37.13%
InternVL2-8B	35.22%	33.51%	32.56%	37.04%	42.01%	34.96%	35.64%	37.33%	39.46%	40.18%	32.72%	41.42%	34.21%	33.66%
InternVL2.5-4B	32.46%	32.43%	31.18%	40.12%	44.38%	22.93%	31.68%	36.34%	44.32%	38.11%	38.89%	36.69%	27.07%	35.15%
InternVL2-2B	31.69%	27.57%	33.95%	35.19%	34.91%	28.57%	29.21%	34.30%	32.43%	37.18%	31.48%	30.18%	34.96%	34.65%
InternVL2.5-1B	28.44%	24.32%	30.95%	32.10%	33.14%	24.06%	25.74%	30.28%	35.68%	26.79%	39.51%	34.32%	27.82%	25.25%
Qwen2.5-VL-72B	36.91%	38.92%	33.49%	38.89%	45.56%	37.59%	32.67%	42.27%	47.57%	40.42%	51.23%	43.20%	37.22%	40.10%
Qwen2-VL-72B	32.67%	37.30%	27.71%	35.19%	37.87%	33.83%	31.19%	38.25%	44.86%	33.49%	43.21%	45.56%	34.96%	36.63%
GLM-4V-9B	31.76%	30.81%	30.02%	30.86%	37.87%	34.59%	28.22%	32.25%	38.92%	38.34%	25.93%	34.32%	25.94%	24.75%
Qwen2.5-VL-7B	30.28%	34.59%	24.25%	30.86%	34.32%	34.21%	30.20%	36.49%	37.84%	34.64%	45.06%	34.91%	38.72%	30.69%
Qwen2-VL-7B	27.73%	27.03%	27.48%	33.95%	33.73%	21.05%	27.72%	31.69%	29.73%	32.10%	37.04%	30.18%	30.83%	30.69%
Qwen2.5-VL-3B	24.28%	24.32%	22.17%	32.72%	26.04%	23.68%	21.29%	32.46%	34.05%	29.79%	40.12%	36.69%	33.08%	26.24%
LongVA	28.09%	30.27%	31.41%	27.16%	28.99%	22.93%	25.74%	32.11%	31.89%	29.10%	30.86%	37.87%	35.71%	30.20%
MiniCPM-V 2.6	28.02%	33.51%	26.79%	27.78%	24.85%	25.19%	32.18%	35.92%	42.16%	34.87%	35.19%	33.73%	38.72%	31.19%
DeepSeek-VL-7B	23.71%	21.08%	23.09%	27.78%	29.59%	22.18%	21.29%	26.11%	25.95%	29.33%	25.93%	30.18%	25.19%	17.33%
Phi-3.5-vision-instruct	21.24%	22.70%	18.94%	18.52%	25.44%	25.19%	18.32%	24.14%	21.62%	25.17%	25.93%	22.49%	24.44%	23.76%
<b>V-Oracle (Qwen2-7B)</b>	<b>81.65%</b>	<b>74.05%</b>	<b>79.21%</b>	<b>85.19%</b>	<b>85.90%</b>	<b>80.45%</b>	<b>89.11%</b>	<b>87.58%</b>	<b>82.16%</b>	<b>84.53%</b>	<b>92.59%</b>	<b>91.72%</b>	<b>87.97%</b>	<b>91.09%</b>

**plexity.** **OBSD** uses a diffusion-based image generation approach, where the input is an oracle bone script image, and the output is an image of a modern Chinese character. To interpret the generated character, an additional **OCR** model is required, introducing

extra complexity to the pipeline. In contrast, **V-Oracle** directly takes an oracle bone script image as input and outputs the corresponding Chinese character in text format, eliminating the need for auxiliary models. This integrated design simplifies the overall workflow and en-

Table 11: The performance of different LMMs on Oracle-Bench from the perspective of character formation principles. The best results are highlighted in **bold**, **blue** for the best closed-source model, and **green** for the best open-source model. (Pic: Pictograms, Ide: Ideograms, Ind: Indicatives).

Model	Decipherment				Interpretation			
	Acc (↑)	Pic (↑)	Ide (↑)	Ind (↑)	Acc (↑)	Pic (↑)	Ide (↑)	Ind (↑)
<i>Closed-source</i>								
GPT-4o	41.00%	67.48%	36.48%	65.91%	44.32%	60.16%	41.63%	59.09%
GPT-4V	29.43%	53.66%	25.46%	50.00%	39.52%	62.60%	36.48%	48.86%
<i>Open-source</i>								
InternVL2.5-78B	43.47%	65.85%	40.30%	55.68%	47.71%	69.92%	44.44%	61.36%
InternVL2-76B	42.48%	61.79%	39.39%	57.95%	44.18%	62.60%	41.79%	51.14%
InternVL2-26B	41.78%	70.73%	37.48%	60.23%	40.08%	57.72%	37.98%	44.32%
InternVL2.5-8B	35.99%	56.10%	33.17%	46.59%	41.78%	55.28%	39.47%	54.55%
InternVL2-8B	35.22%	52.85%	32.17%	52.27%	37.33%	56.91%	35.41%	36.36%
InternVL2.5-4B	32.46%	45.53%	29.68%	52.27%	36.34%	47.97%	34.25%	48.86%
InternVL2-2B	31.69%	44.72%	29.52%	43.18%	34.30%	41.46%	33.00%	42.05%
InternVL2.5-1B	28.44%	30.89%	27.94%	31.82%	30.28%	40.65%	28.36%	42.05%
Qwen2.5-VL-72B	36.91%	58.54%	33.67%	51.14%	42.27%	60.16%	40.13%	46.59%
Qwen2-VL-72B	32.67%	59.35%	29.35%	40.91%	38.25%	53.66%	36.32%	43.18%
Qwen2.5-VL-7B	30.28%	56.91%	27.03%	37.50%	36.49%	55.28%	34.41%	38.64%
Qwen2-VL-7B	27.73%	47.15%	24.79%	40.91%	31.69%	37.40%	30.85%	35.23%
Qwen2.5-VL-3B	24.28%	39.02%	22.22%	31.82%	32.46%	42.28%	31.26%	35.23%
GLM-4V-9B	31.76%	55.28%	28.94%	37.50%	32.25%	46.34%	30.85%	31.82%
LongVA	28.09%	55.28%	24.38%	40.91%	32.11%	39.84%	31.51%	29.55%
MiniCPM-V 2.6	28.02%	35.77%	26.53%	37.50%	35.92%	47.97%	34.74%	35.23%
DeepSeek-VL-7B	23.71%	33.33%	23.13%	18.18%	26.11%	30.08%	25.29%	31.82%
Phi-3.5-vision-instruct	21.24%	28.46%	19.98%	28.41%	24.14%	22.76%	23.71%	31.82%
<b>V-Oracle (Qwen2-7B)</b>	<b>81.72%</b>	<b>90.24%</b>	<b>80.35%</b>	<b>88.64%</b>	<b>87.58%</b>	<b>92.68%</b>	<b>87.06%</b>	<b>87.50%</b>

hances model efficiency.

- **Flexible Outputs vs. Fixed Input-Output Formats.** As an image generation model, **OBSD** has fixed input-output formats, where the input must be an image and the output is also an image. This limits its ability to handle diverse use cases. In contrast, **V-Oracle** integrates a decoder based on a large language model (LLM), which not only provides textual outputs but also supports interactive dialogue-based responses. This flexibility allows **V-Oracle** to serve a wider range of applications, including providing explanations or answering queries in natural language, making it suitable for users from various domains.

**Results** Although **OBSD** has not fully disclosed its training process or the specifics of its test set, making a direct comparison of the results infeasible, we reference the accuracy reported in the **OBSD** paper, which stands at 41%. Given that the test set sources for both methods are similar, **V-Oracle** demonstrates significant improvements, achieving an accuracy of 81.65% on in-domain decipherment tasks and 59.8% on out-of-domain

tasks. These results highlight the superior performance of **V-Oracle** in oracle bone script decipherment.

## E.2 Differences from Other Studies

There are currently no works in the field of large models that focus on oracle bone scripts, nor does any existing research attempt to decipher oracle bone scripts using such models. Previous studies mainly address tasks such as classification (Wang and Deng, 2024; Wang et al., 2024d,a), retrieval (Han et al., 2020b), and denoising of oracle bone scripts (Wang et al., 2022; Qiao et al., 2024b). With the rise of large multimodal models (LMMs), our work is the first to fill this gap, utilizing LMMs to tackle the challenge of oracle bone script decipherment.

## E.3 Comparative Analysis of Oracle-Bench and OBI-Bench

**OBI-Bench** (Chen et al., 2024b) evaluates the fine-grained perception and cognition abilities of LMMs in the whole-process **OBI** tasks, which include recognition, rejoining, classification, retrieval, and deciphering. In contrast, our **Oracle-Bench** focuses

on exploring the ability to decipher OBS based on the logic of pictographic meanings. A multidimensional comparison of the two is presented in Table 12.

Table 12: Benchmark Feature Comparison

Benchmark	Samples	Subfields Classification	Pictographic Annotations
Oracle-Bench (Ours)	2,834	✓	✓
OBI-Bench (Decipher)	140	-	-

## F Detailed Related Work

The rapid development of large language models (LLMs), such as GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2023), Qwen (Bai et al., 2023), and Deepseek (Liu et al., 2024a), has significantly advanced natural language processing (Ying et al., 2024; Song et al., 2024; Lei et al., 2023; Zhang et al., 2023; ?), enabling breakthroughs in instruction following (Zhou et al., 2023; Dong et al., 2024), retrieval (Jin et al., 2025b; Li et al., 2024d) and search (Li et al., 2025; Jin et al., 2025a). Building upon these foundations, large multimodal models (LMMs) have emerged, integrating powerful language models with visual encoders to achieve impressive performance across a wide range of domains. Both open-source (Chen et al., 2024a; Wang et al., 2024b; GLM et al., 2024; Abdin et al., 2024; Yao et al., 2024; Zhang et al., 2024; Lu et al., 2024; Team, 2025) and closed-source (OpenAI, 2024, 2023) LMMs have demonstrated strong capabilities. For instance, the LLaVA (Liu et al., 2024d) series standardizes visual instruction tuning via a linear projector, while InternVL (Chen et al., 2024a) employs a large vision encoder and QFormer for high-quality visual integration. LLaVA-NeXT (Liu et al., 2024c) introduces the "AnyRes" technique for flexible image resolution, and LLaVA-OneVision (Li et al., 2024b) further excels in video and multi-image tasks.

These advances have enabled LMMs to tackle fundamental tasks in diverse domains, including mathematical reasoning (Lu et al., 2023; Qiao et al., 2024a; Zhuang et al., 2024), medical (Li et al., 2024c; Saab et al., 2024), and OCR (Li et al., 2024e; Hu et al., 2024; Wei et al., 2024), demonstrating remarkable adaptability and potential. However, in the domain of oracle bone scripts, significant challenges persist. The limited availability of data and the substantial differences between pictographic oracle bone scripts and modern visual data present unique obstacles. To address these

challenges, we leverage the pictographic nature of oracle bone scripts, employ extensive data augmentation, and model the decipherment process to propose V-Oracle, bridging this critical gap and offering a novel approach to oracle bone research.

## G Broaden Impact

**Cultural Value.** Oracle bone script, as one of the earliest forms of Chinese characters, carries significant historical value and serves as an important way to explore the development of human civilization and ancient social structures. It records various aspects of life during the Shang dynasty, including politics, economy, religion, and daily activities, providing invaluable first-hand materials for studying the mechanisms and evolution of ancient Chinese society. As a cornerstone of Chinese civilization and a part of the world’s cultural heritage, it reflects the intelligence and creativity of ancient humanity, with its importance extending far beyond regional and national boundaries.

With the rapid development of LMMs, our work has significantly improved the efficiency of oracle bone decipherment, making it easier for the public to access and understand oracle bones. Additionally, through OOD experiments, we have demonstrated the practical value of **V-Oracle**. As a deciphering tool, **V-Oracle** provides strong support for archaeologists, helping them decode oracle bones more effectively. This combination of technology and humanities not only advances oracle bone research but also offers a new direction for the digital preservation and global dissemination of cultural heritage.

**Educational Value.** The study of oracle bone script is not only a specialized academic endeavor but also a rich educational resource that connects history, culture, and technology. By integrating advanced large model technologies, our work introduces a novel way to make the study of oracle bones more accessible and engaging. This approach bridges the gap between traditional cultural studies and modern audiences, especially younger generations, by providing interactive tools and intuitive platforms.

Through the use of frameworks like **V-Oracle**, students and educators can gain deeper insights into the complexities of ancient scripts and their historical contexts. These tools not only simplify the learning process but also spark curiosity about ancient cultures and inspire a new generation of


scholars to contribute to the preservation and study of cultural heritage. By combining history with cutting-edge technology, our work offers an innovative model for educational practices, fostering a deeper appreciation for cultural heritage in the digital age.

**Expanding Applications of LMMs. V-Oracle** offers new possibilities for deciphering not only oracle bone script but also other ancient languages, such as cuneiform, Mayan hieroglyphs, and ancient Egyptian texts. By leveraging the flexibility and adaptability of large multimodal models (LMMs), researchers can extend their work to multiple linguistic systems, providing innovative methods for uncovering the stories embedded in historical records. This cross-linguistic capability paves the way for new insights into human history and global cultural exchange.

At the same time, similar to recent advancements in fields like medicine (Li et al., 2024c), the rapid development of LMMs highlights their potential across a wide range of disciplines. From linguistics to cultural studies, and from science and technology to societal applications, the cross-domain capabilities of these models are continuously reshaping traditional research methods and fostering deeper interdisciplinary collaboration. In the future, the diverse applications of LMMs may redefine conventional fields of study and infuse innovation into various academic disciplines.



**Deciphering Oracle Bone Script "光 (Light)" with V-Oracle**



**Prompt:**  
Now, we need you to answer a multiple-choice question about Oracle Bone Script. Regarding the Oracle Bone Script image, please determine which modern Chinese character the Oracle Bone Script in the image most likely corresponds to from the following options:

**Options:**  
A. 娘 (Mother) B. 黍 (Millet) C. 函 (Box) D. 光 (Light)

Please first provide your reasoning in the <Thought Process> section, and finally give your answer choice in the <Answer> section, formatted as follows:  
<Thought Process>:<Your Reasoning>  
<Answer>:<Option>

---

**Response from V-Oracle:**  
<Thought process>: It can be observed that the Oracle Bone Script shown in the image symbolizes a person holding a torch, representing a person holding a source of illumination. This corresponds to the modern Chinese character '光'(Light), and the original meaning of the character '光' is brightness or illumination. The image in the Oracle Bone Script vividly depicts the source of light—a torch—through the combination of 'person' and 'fire', which aligns with the meaning of the modern character '光'. Therefore, the final answer is 'D'.



<Answer>: D. 光 (Light) 

Figure 11: A successful decipherment case by V-Oracle.

**Deciphering Oracle Bone Script "山 (Mountain)" with V-Oracle**




**Prompt:**  
Now, we need you to answer a multiple-choice question about Oracle Bone Script. Regarding the Oracle Bone Script image, please determine which modern Chinese character the Oracle Bone Script in the image most likely corresponds to from the following options:

**Options:**  
A. 山 (Mountain) B. 雨 (Rain) C. 丘 (Hill) D. 城 (City)

Please first provide your reasoning in the <Thought Process> section, and finally give your answer choice in the <Answer> section, formatted as follows:  
<Thought Process>:<Your Reasoning>  
<Answer>:<Option>

---

**Response from V-Oracle:**  
<Thought process>: It can be observed that the Oracle Bone Script shown in the image symbolizes connected mountain peaks, representing two peaks joined together, which corresponds to the modern Chinese character '丘' (Hill). The character '丘' itself means a small hill or mound, and the image in the Oracle Bone Script vividly depicts this scene of two connected hills, which is a concrete pictorial representation of '丘'. Therefore, the final answer is 'C'.

<Answer>: C. 丘 (Hill) 




	
山 (Mountain)	丘 (Hill)

Figure 12: A failed decipherment case by V-Oracle.

### Comparison of V-Oracle and Open-Source Model in OOD Setting



饮  
(Drink)

**Prompt:**

Now, we need you to answer a multiple-choice question about Oracle Bone Script. Regarding the Oracle Bone Script image, please determine which modern Chinese character the Oracle Bone Script in the image most likely corresponds to from the following options:


Options:  
A. 稷 (Grain) B. 饮 (Drink) C. 街 (Street) D. 吃 (Eat)

Please first provide your reasoning in the <Thought Process> section, and finally give your answer choice in the <Answer> section, formatted as follows:  
<Thought Process>:<Your Reasoning>  
<Answer>:<Option>

---

**Response from InternVL2-76B:**

<Thought process>: The oracle bone script shown in the image resembles the character '稷', which represents grains or crops. The shapes of the oracle bone scripts in the other options do not match the image. Therefore, the oracle bone script shown in this image most likely corresponds to the modern Chinese character '稷'.

<Answer>: A 

**Response from V-Oracle:**

<Thought process>: It can be observed that the oracle bone script shown in the image symbolizes a person raising a cup of wine for a ritual, indicating the act of sprinkling wine towards the sky as part of a sacrificial ceremony. This corresponds to the modern Chinese character '饮' (drink). The oracle bone script perfectly depicts this action and scene. Therefore, it can be inferred that this oracle bone script corresponds to the modern Chinese character '饮'. Hence, the final answer is B.


<Answer>: B. 

Figure 13: Comparison of V-Oracle and an open-source model in the OOD setting, highlighting how V-Oracle infers correctness through pictographic understanding, while the open-source model relies solely on direct character-to-modern-Chinese mapping.

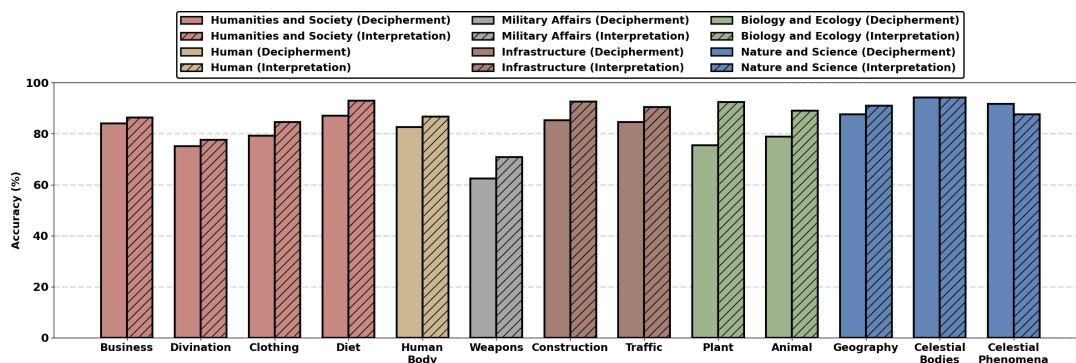


Figure 14: Detailed performance of V-Oracle across 13 knowledge second-level nodes.

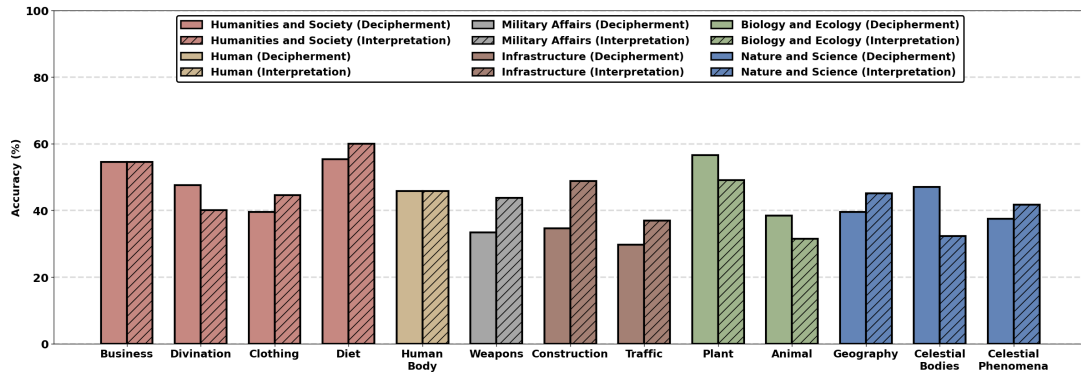


Figure 15: Detailed performance of GPT-4o across 13 knowledge second-level nodes.

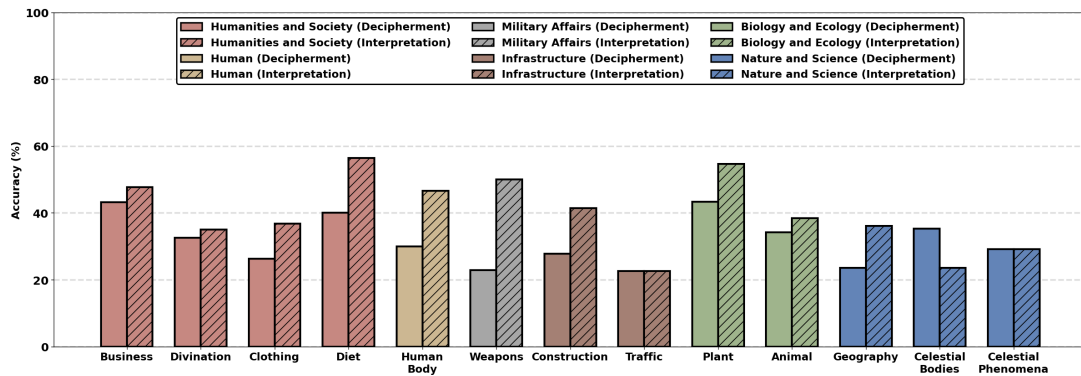


Figure 16: Detailed performance of GPT-4V across 13 knowledge second-level nodes.

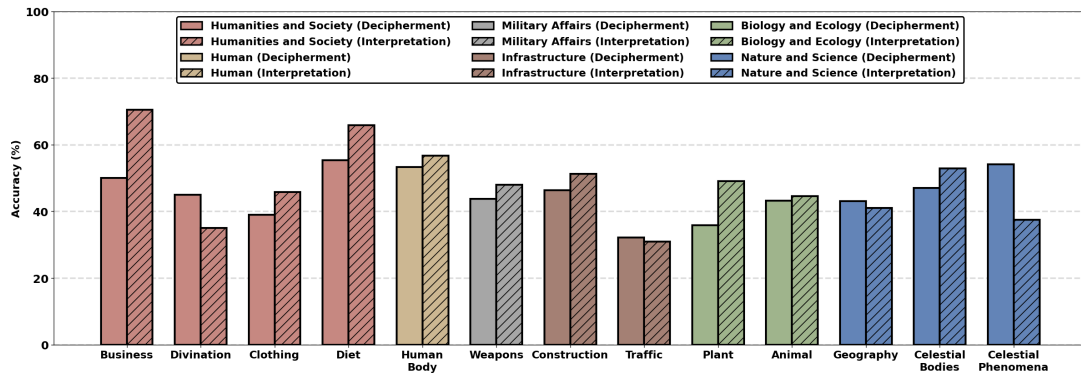


Figure 17: Detailed performance of InternVL2.5-78B across 13 knowledge second-level nodes.

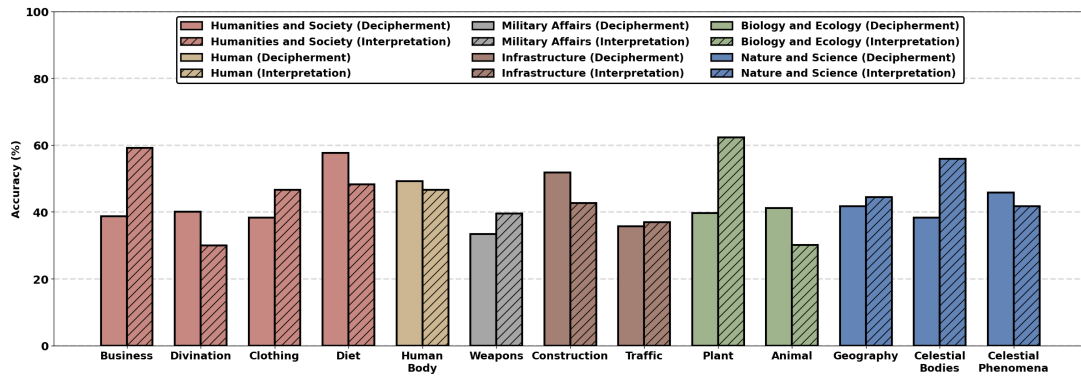


Figure 18: Detailed performance of InternVL2-Llama3-76B across 13 knowledge second-level nodes.

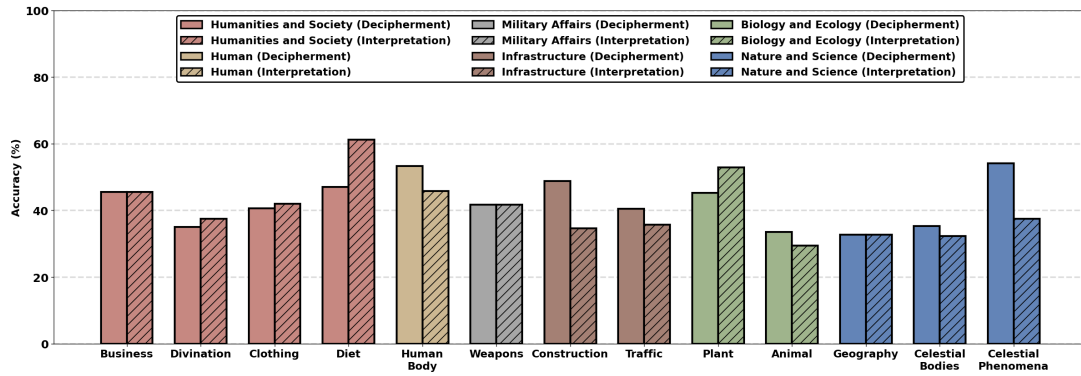


Figure 19: Detailed performance of InternVL2-26B across 13 knowledge second-level nodes.

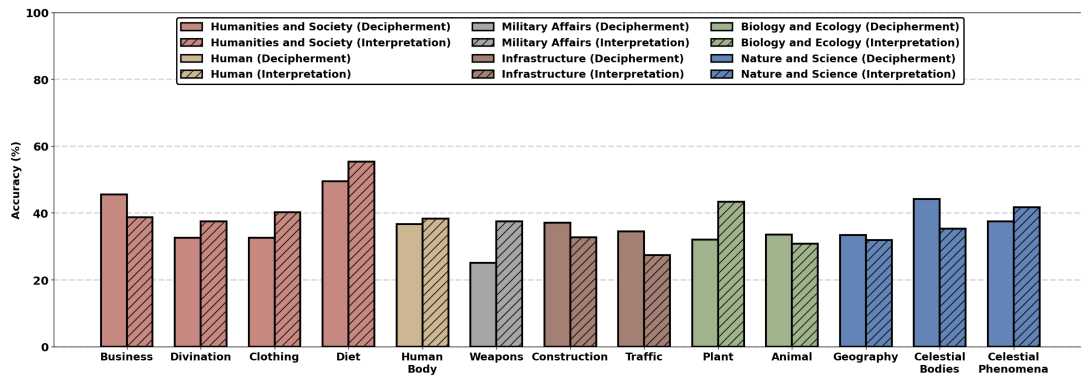


Figure 20: Detailed performance of InternVL2-8B across 13 knowledge second-level nodes.

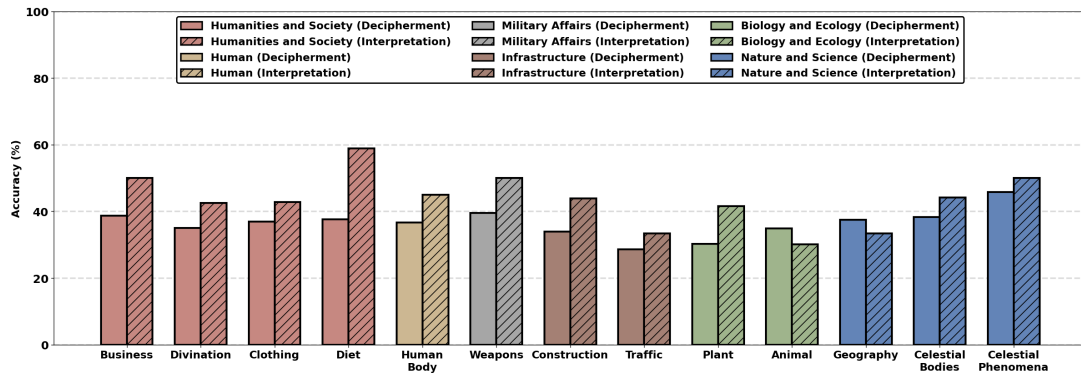


Figure 21: Detailed performance of InternVL2.5-8B across 13 knowledge second-level nodes.

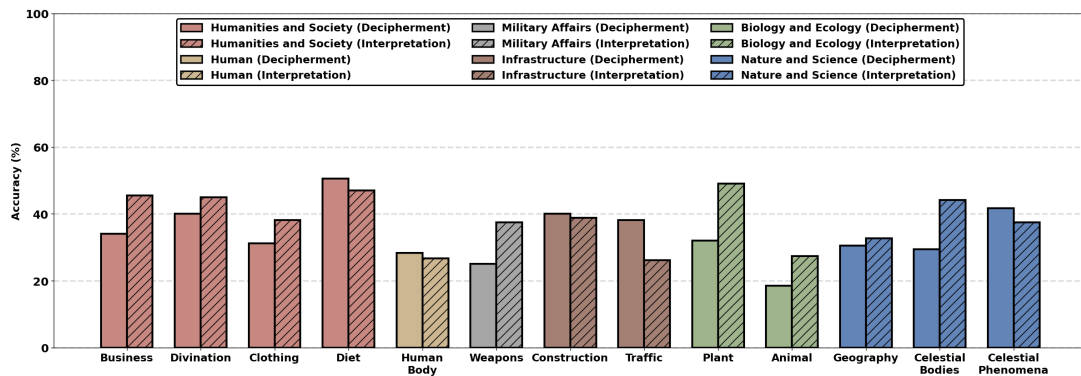


Figure 22: Detailed performance of InternVL2.5-4B across 13 knowledge second-level nodes.



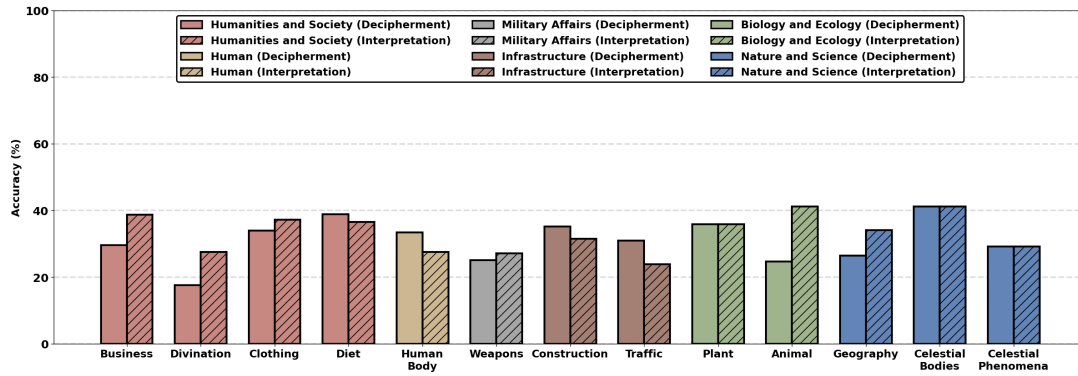


Figure 23: Detailed performance of InternVL2-2B across 13 knowledge second-level nodes.

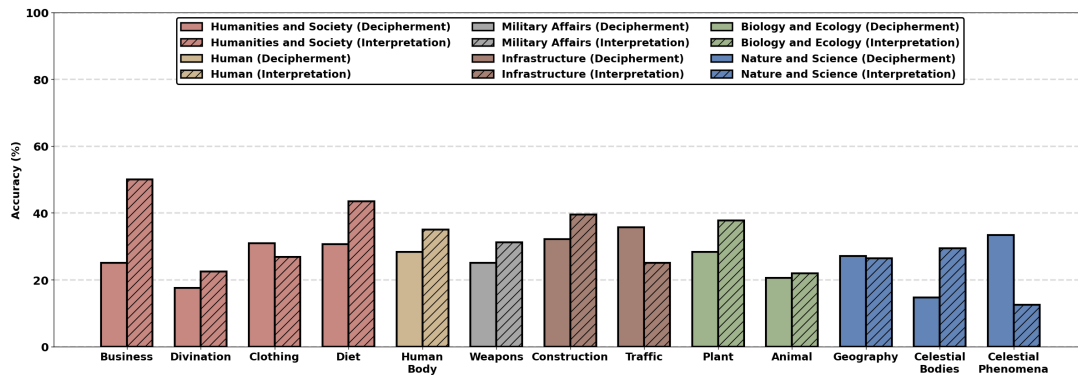


Figure 24: Detailed performance of InternVL2.5-1B across 13 knowledge second-level nodes.

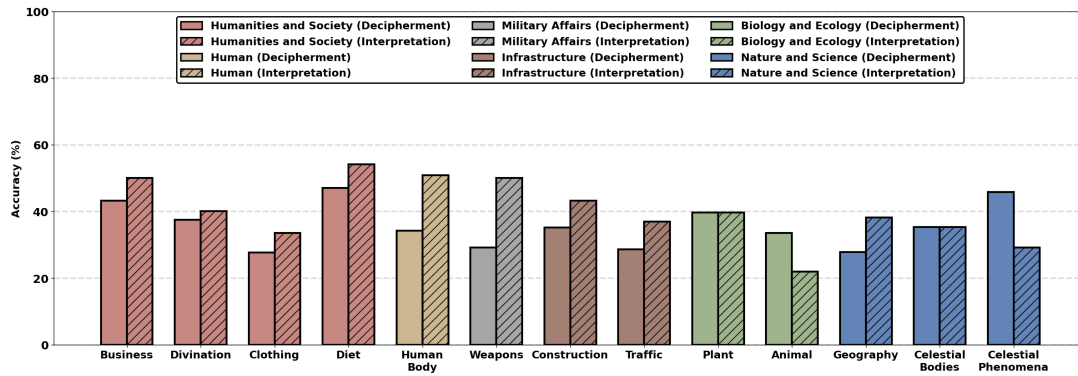


Figure 25: Detailed performance of Qwen2-VL-72B across 13 knowledge second-level nodes.

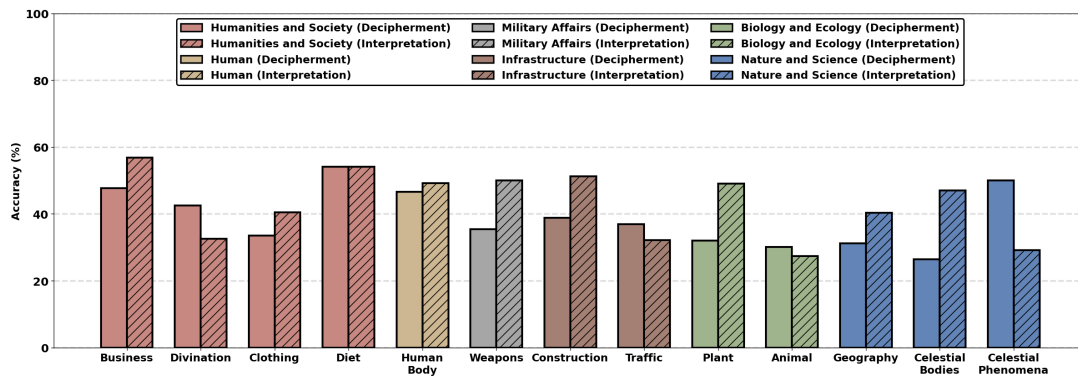


Figure 26: Detailed performance of Qwen2.5-VL-72B across 13 knowledge second-level nodes.

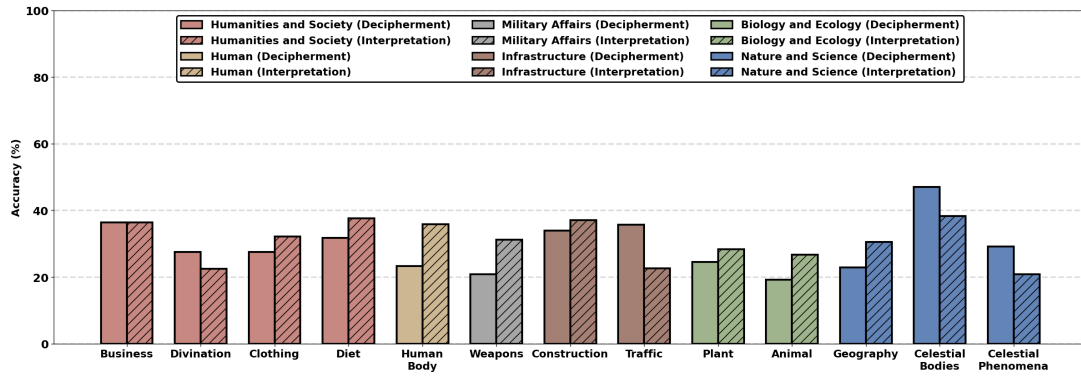


Figure 27: Detailed performance of Qwen2-VL-7B across 13 knowledge second-level nodes.

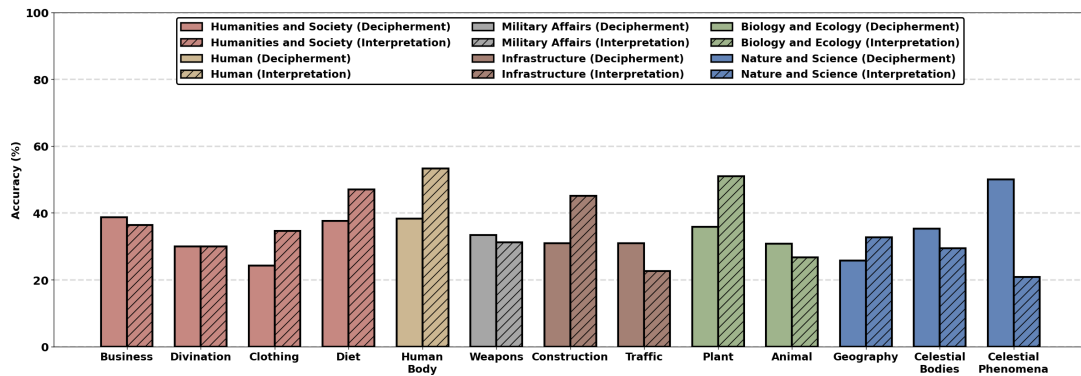


Figure 28: Detailed performance of Qwen2.5-VL-7B across 13 knowledge second-level nodes.

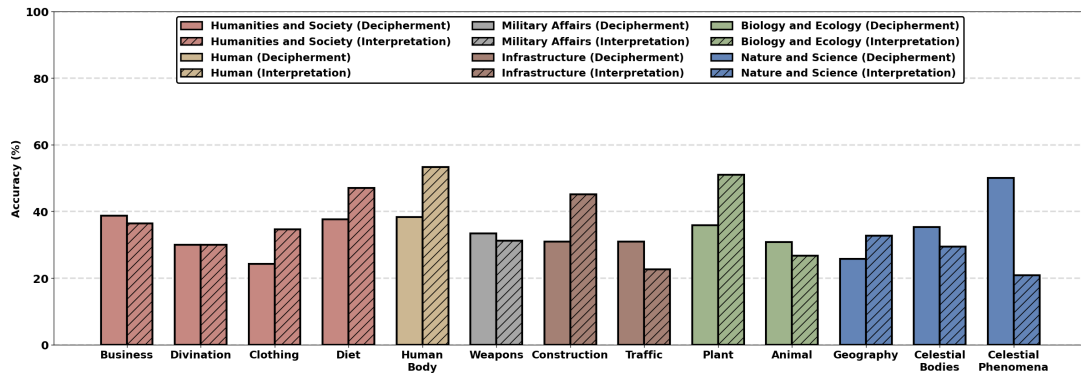


Figure 29: Detailed performance of Qwen2.5-VL-3B across 13 knowledge second-level nodes.

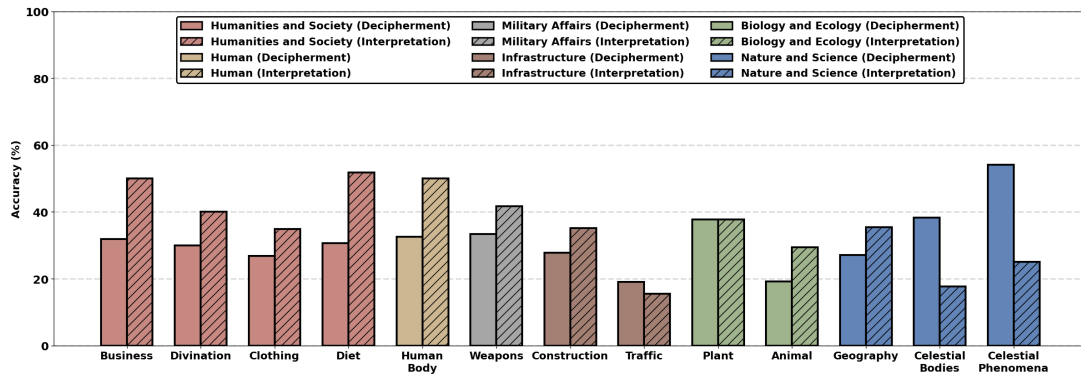


Figure 30: Detailed performance of MiniCPM-V 2.6 across 13 knowledge second-level nodes.

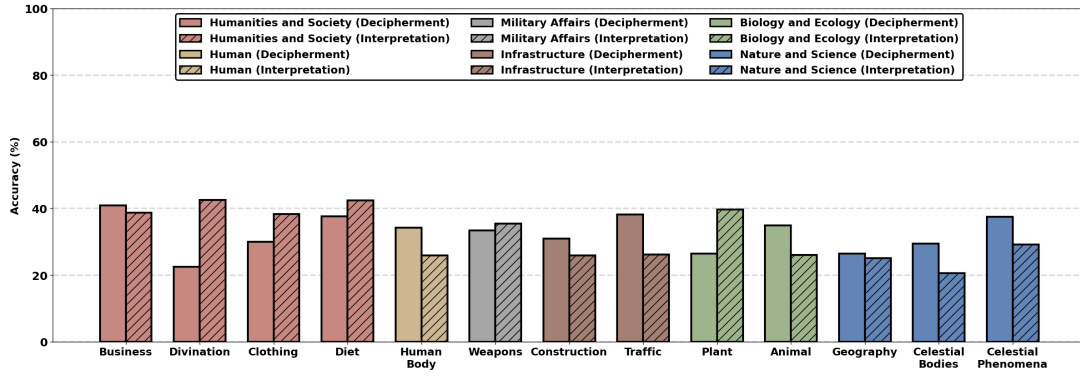


Figure 31: Detailed performance of GLM-4V-9B across 13 knowledge second-level nodes.

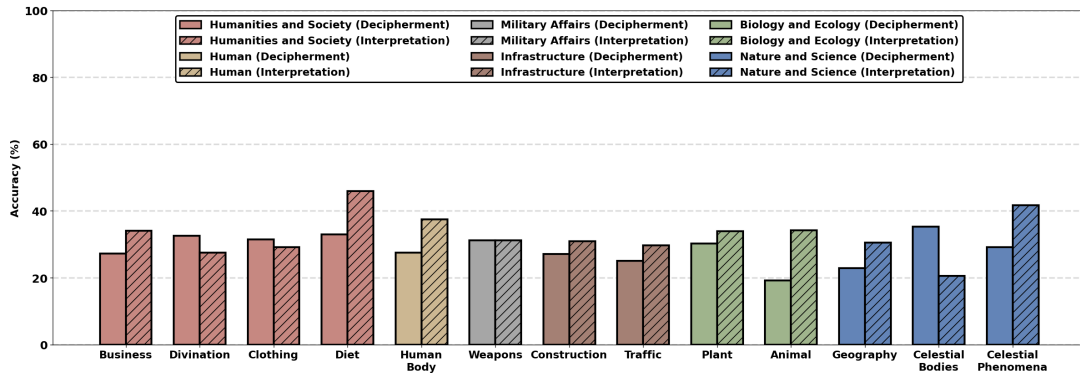


Figure 32: Detailed performance of LongVA across 13 knowledge second-level nodes.

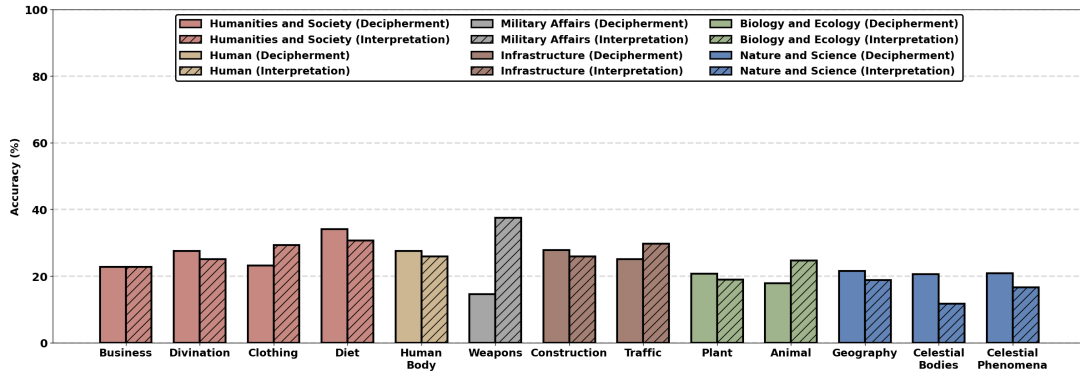


Figure 33: Detailed performance of DeepSeek-VL-7B across 13 knowledge second-level nodes.

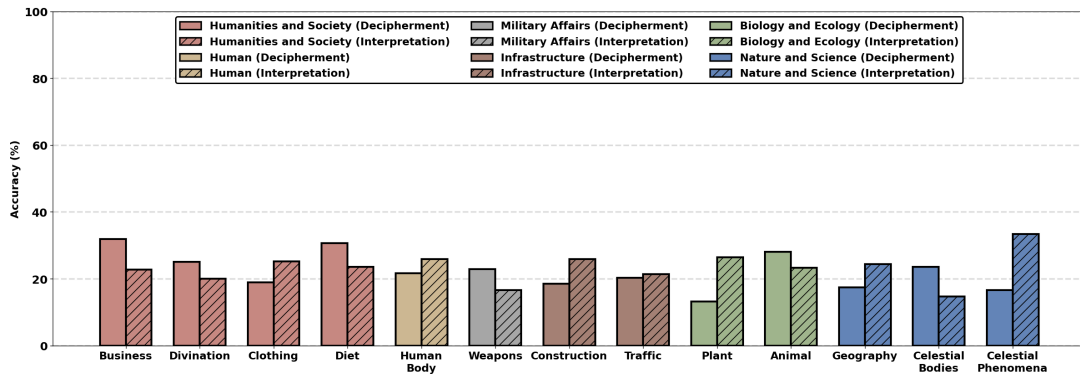


Figure 34: Detailed performance of Phi-3.5-vision-instruct across 13 knowledge second-level nodes.