DongbaMIE: A Multimodal Information Extraction Dataset for Evaluating Semantic Understanding of Dongba Pictograms

Xiaojun Bi[*,✉,1,2], Shuo Li[*,3], Junyao Xing[3], Ziyue Wang[4], Fuwen Luo[4], Weizheng Qiao[1,2], Lu Han[1,2], Ziwei Sun[1,2], Peng Li[5], Yang Liu[4,5]

[1]College of Information and Engineering, Minzu University of China, Beijing, China
[2] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing, China
[3] College of Information and Communication Engineering, Harbin Engineering University, Harbin, China
[4] Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China
[5] Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
{bixiaojun, thinklis, 595587572}@hrbeu.edu.cn,
{wangziyue22, lfw23}@mails.tsinghua.edu.cn
{qiaoweizheng, 22400208, 22400215}@muc.edu.cn
lipeng@air.tsinghua.edu.cn, liuyang2011@tsinghua.edu.cn

Abstract

Dongba pictographic is the only pictographic script still in use in the world. Its pictorial ideographic features carry rich cultural and contextual information. However, due to the lack of relevant datasets, research on semantic understanding of Dongba hieroglyphs has progressed slowly. To this end, we constructed DongbaMIE - the first dataset focusing on multimodal information extraction of Dongba pictograms. The dataset consists of images of Dongba hieroglyphic characters and their corresponding semantic annotations in Chinese. It contains 23,530 sentence-level and 2,539 paragraph-level high-quality text-image pairs. The annotations cover four semantic dimensions: object, action, relation and attribute. Systematic evaluation of mainstream multimodal large language models shows that the models are difficult to perform information extraction of Dongba hieroglyphs efficiently under zero-shot and few-shot learning. Although supervised fine-tuning can improve the performance, accurate extraction of complex semantics is still a great challenge at present.[1]

1 Introduction

Dongba pictographic script is currently mainly used by the Naxi ethnic group in southwest China. As an important part of the cultural heritage of the Naxi ethnic group in China (Luo et al., 2023b; Xu, 2023), ancient books written in Dongba pictographic script have been officially listed in the International Memory of the World Register by the United Nations Educational, Scientific and Cultural Organization (UNESCO). Figure 1 shows an image of Dongba pictographic script, which presents content composed of exquisite handwriting. It is recorded on local paper and mainly involves Dongba religion, Naxi history and lifestyle. However, there are severe challenges in the digital preservation and processing of Dongba pictographic script (Wang et al., 2011). First, since there are very few people who understand Dongba pictographic script, only a few Naxi priests ("Dongba") can read this script today. This makes the relevant modern language annotation resources extremely scarce. Secondly, Dongba pictographic script lacks a standardized encoding system, and its grammatical structure is also significantly different from that of modern languages (Bi and Luo, 2024). This makes it impossible to process it like regular text.

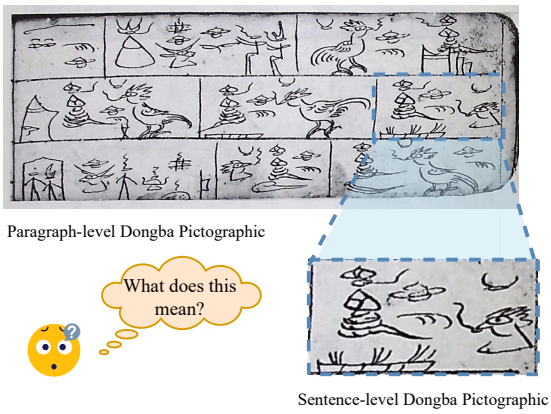


Figure 1: An example of Dongba pictograms from the DongbaMIE dataset.

[*]Equal contribution.
[✉]Corresponding authors.
[1]Our code and data is available at https://github.com/thinklis/DongbaMIE.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 976–990
November 4-9, 2025

These factors make it difficult to apply traditional natural language processing methods to the processing of Dongba pictographic scripts, and further aggravate the severe shortage of modern annotation corpus. Therefore, it is particularly important to construct a multimodal semantic understanding dataset based on Dongba pictograms, which will provide valuable resources for research and application in this field.

Recent advances in deep learning have provided new opportunities for endangered language processing (Sommerschield et al., 2023; Lu, 2024; Zhang et al., 2024, 2022). However, the understanding of contextual semantic information in Dongba pictograms has not been further studied. In addition, Dongba pictograms have unique linguistic phenomena. This includes the omission of pictographs in contextual sentences and the presence of polysemous characters with multiple meanings. These characteristics further increase the difficulty of deciphering and preserving these endangered texts through computation.

To improve the semantic understanding of Dongba hieroglyphs, we introduce a novel multimodal information extraction dataset, DongbaMIE, which is derived from "The Annotated Collection of Naxi Dongba Manuscripts" (He and He, 1999). DongbaMIE contains 23,530 sentence-level and 2,539 paragraph-level images of precisely scanned manuscripts. It is also annotated with object, action, relationship and attribute information from its Chinese translation text. We evaluated multimodal large language models (MLLMs) at DongbaMIE. This includes zero-shot and few-shot performance of proprietary models such as GPT-4o (Hurst et al., 2024), and supervised fine-tuning (SFT) performance of open-source models such as LLaVA-NeXT (Liu et al., 2024b) and Qwen2-VL (Wang et al., 2024). Experiments show that current MLLMs struggle with this task. For example, the object extraction F1 of GPT-4o is only 1.60 under zero-shot. Despite the performance improvement of SFT, MLLMs show significant shortcomings in extracting complex semantics, especially in extracting relations and attributes. Our contributions are as follows:

- We introduce DongbaMIE, the first multimodal information extraction dataset for Dongba pictograms, which provides richly annotated image-text pairs covering four key semantic dimensions.
- We systematically evaluate the mainstream MLLMs on DongbaMIE, assessing their performance in different settings such as zero-shot, few-shot, and SFT.
- We find that MLLMs have significant limitations in performing multimodal information extraction tasks for Dongba pictograms, especially in complex semantic understanding.

2 Related Work

2.1 Processing of Ancient Endangered Languages

In recent years, the digitization of ancient endangered languages has gradually become an important research direction in NLP (Anderson et al., 2023; Pavlopoulos et al., 2024). These languages face a serious data scarcity problem. Their writing systems usually have diverse and complex morphological features. This makes it difficult to directly use traditional text processing methods (Ignat et al., 2022; Buoy et al., 2023; Gunna et al., 2021; Tan et al., 2020; Deng and Liu, 2018). With the development of technology, related research has also experienced a transition from basic digitization to more advanced language analysis tasks (Sommerschield et al., 2023). This includes semantic and sentiment analysis (Yoo et al., 2022; Sahala et al., 2020; Pavlopoulos et al., 2022), translation (Kang et al., 2021; Yousef et al., 2022; Jin et al., 2023), and decryption (Luo et al., 2021; Daggumati and Revesz, 2018). The EvaLatin Challenge uses Perseus and LASLA corpora for Latin part-of-speech tagging and word form restoration (Sprugnoli et al., 2024, 2022). Yoo et al. (2022) proposed a Transformer-based analysis of Chinese historical documents. Jin et al. (2023) performed morphological and semantic evaluation on machine translation of ancient Chinese. However, the above research is still a text-centric approach. This makes it difficult to use it in ideographic writing systems with strong semantics.

In addition, Sahala et al. (2020) proposed a character-level sequence-to-sequence model, which achieved the automatic transcription of Akkadian transliterated text for the first time. Gordin et al. (2024) proposed an OCR pipeline for digitizing cuneiform transliterated data. At the same time, De Cao et al. (2024) proposed a method for translating ancient Egyptian hieroglyphs based on the Transformer model by combining the characteristics of speech and semantics. These methods

mainly use the characteristics of speech and semantics to associate with the target language. Dongba hieroglyphs lack similar linguistic tool support due to their limited application and separation from the modern language system.

Existing research on Dongba pictograms has primarily focused on character-level processing and statistical machine translation. In the former, efforts have concentrated on tasks such as character detection to localize and classify characters in images (Ma et al., 2024), and the recognition of a limited number of isolated handwritten characters (Luo et al., 2023b). In the latter, there have been attempts to build statistical translation models based on dependency structures (Gao et al., 2017, 2018). However, these approaches face significant challenges. Due to the lack of a unified encoding, fixed grammatical structures, and a standard benchmark dataset, existing methods struggle to capture contextual semantic information, leaving the broader language analysis of Dongba pictograms in its nascent stages. To address these limitations, we introduce the DongbaMIE dataset, shifting the focus from character-level analysis to sentence- and paragraph-level multimodal information extraction. By providing fine-grained annotations across the four dimensions of objects, actions, relations, and attributes that link pictographs to structured semantics, our work aims to facilitate the development and evaluation of models like LLMs on complex semantic reasoning tasks, addressing a notable gap in existing resources.

2.2 Multimodal Semantic Analysis of Pictographic Characters

In recent years, multimodal information extraction techniques have shown great potential in processing vision-language interaction tasks (Liu et al., 2019; Sun et al., 2024; Luo et al., 2023a; Diao et al., 2025). Existing studies have made significant progress in image-text alignment and understanding of historical documents and other pictographic symbols (Yang et al., 2023; El Bahi, 2024; Carlson et al., 2024) used contrastively trained visual encoders to model OCR as a character-level image retrieval problem. This achieved more accurate OCR results in historical documents. However, these methods mainly target standardized writing systems with clear encoding schemes. Recently, some studies have processed ideograms through novel multimodal methods. The OBI-BENCH benchmark (Chen et al., 2025) evaluates large multimodal models on oracle bone script tasks (recognition, reconstruction, classification, retrieval, and interpretation). And LogogramNLP (Chen et al., 2024) explores the learning of visual representations of ancient ideograms. These studies show that cultural heritage data can be unlocked by using multimodal processing methods.

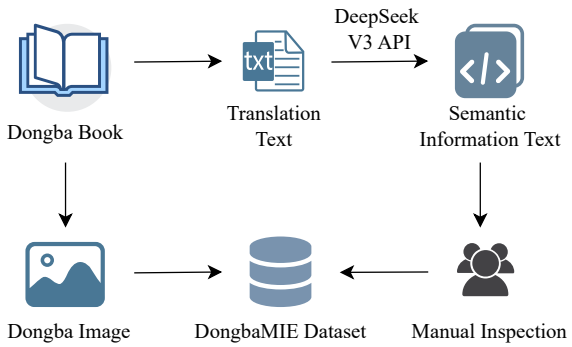


Figure 2: The entire process of establishing the DongbaMIE dataset.

While multimodal approaches have made progress in the identification of other historical scripts, they still struggle with the unique Dongba hieroglyphs. The lack of a standardized Dongba hieroglyphic dataset severely hampers the development and comparison of custom models.

3 Constructing DongbaMIE Dataset

3.1 Data Collection

The DongbaMIE dataset is derived from the 100-volume "An Annotated Collection of Naxi Dongba Manuscripts" (He and He, 1999), a foundational compilation of expertly translated and verified Dongba hieroglyphic manuscripts previously available only in print. For this work, we focus on the initial ten volumes. This scope was deliberately chosen to ensure a balanced representation of key Dongba linguistic phenomena and pictographic features. While the subsequent ninety volumes cover more specialized thematic content, such as regional rituals, their fundamental semantic and syntactic structures are largely consistent with our selected volumes. Therefore, this initial collection is sufficient for developing and evaluating foundational models for multimodal information extraction. We obtained scanning authorization through the university library and worked with professional organizations. We specified a three-phase digitization program. Future work will involve progressively incorporating the remaining volumes to enhance

Split	Sentence	Paragraph	Object		Action		Relation		Attribute	
			Sent	Para	Sent	Para	Sent	Para	Sent	Para
Train	18,824	2,031	65,264	52,011	33,658	24,190	30,165	29,160	19,754	19,085
Dev	2,353	254	7,966	6,413	4,144	2,866	3,720	3,623	2,653	2,528
Test	2,353	254	8,142	6,716	4,199	3,052	3,715	3,974	2,541	2,497
Total	23,530	2,539	81,372	65,140	42,001	30,108	37,600	36,757	24,948	24,110

Table 1: Statistics of the DongbaMIE dataset. "Sent" and "Para" denote sentences and paragraphs, respectively.

the dataset's thematic breadth.

High-Fidelity Imaging and Preprocessing: Images were acquired using professional-grade, high-resolution (e.g., 600 dpi) non-contact book scanners. Through pre-physical flattening, exposure and contrast optimization, and subsequent automatic paging and image enhancement processing, the JPG format image is ensured to retain the original handwriting details and spatial layout to the greatest extent.

Structural Segmentation: Leveraging the unique vertical separators in Dongba pictograms as punctuation, we precisely segmented paragraph-level images into sentence-level images, while also retaining paragraph-level images to support multi-granularity analysis.

Cross-Modal Alignment and Textual Correction: Chinese translations within the images were initially extracted using OCR. However, all textual content is strictly manually proofread word by word. Trained annotators rectified errors by referencing the original manuscripts, ensuring the accuracy of the TXT-formatted text.

To ensure data quality, we established a multi-tiered Quality Assurance mechanism. This mechanism included real-time reviews by a professional team, systematic pre-service training for annotators, and the deep involvement of senior Dongba research experts throughout the process. These experts provided professional guidance, resolved ambiguous cases, and oversaw final quality control, guaranteeing the dataset's scholarly rigor.

3.2 Annotation Scheme

We propose a multidimensional annotation framework which includes objects, relations, actions and attributes. The design of the framework is based on a systematic analysis of grammatical structures observed in Chinese translations of Dongba hieroglyphic. Our analysis reveals consistent correlations in which noun constituents typically denote objects, verb constituents express actions, prepositional structures encode relations, and modifiers describe attributes. Thus, this multidimensional structure effectively captures the core semantic elements of a sentence. Figure 3 provides an example detailing the Dongba hieroglyphs, their Chinese translation, and the corresponding information extraction results. Our multidimensional annotation framework contains four dimensions:

Object: They represent named entities and concepts derived from Chinese translations (e.g., "priest", "wine"). These concepts are the key basic narrative units for understanding the theme.

Action: They represent events and activities, usually expressed as verbs (e.g., "offer a ritual sacrifice", "offer a toast"). Actions constitute the narrative flow and its temporal causality.

Relation: They encode explicit connections and interactions between objects (e.g., "priest-holding-cypress branches"). This dimension is essential for understanding the interactions between entities and maintaining contextual integrity.

Attribute: They specify descriptive features of an object or entity (e.g., "wine-tastes-aromatic", "cypress branches-color-emerald green") that add detail and depth to the semantic representation.

Overall, these four dimensions aim to holistically capture the multilayered semantics of Dongba hieroglyphs. Objects and their attributes constitute the core semantic entities, while actions and relationships describe the dynamic interactions in the narrative. The framework aims to balance the simplicity of annotation with comprehensive semantic coverage, providing a robust and scalable foundation for subsequent structured text analysis.

3.3 Hybrid Annotation Pipeline

To ensure the accuracy and reliability of semantic annotations in DongbaMIE, we implemented a rigorous hybrid annotation process. This process integrates automated pre-annotation using LLM with multi-stage manual verification and quality control by trained annotators. The overall work-

Model	Level		Object			Action			Relation			Attribute		
			P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-4o	sent	0-shot	1.73	1.80	1.60	0.34	0.39	0.32	0.00	0.00	0.00	0.00	0.00	0.00
		1-shot	1.55	1.57	1.45	0.32	0.42	0.32	0.00	0.00	0.00	0.00	0.00	0.00
	para	0-shot	5.91	1.99	2.88	1.07	0.36	0.49	0.00	0.00	0.00	0.00	0.00	0.00
		1-shot	6.85	2.18	3.16	1.28	0.61	0.76	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-2.0	sent	0-shot	1.85	2.03	1.77	1.04	1.07	0.93	0.00	0.00	0.00	0.00	0.00	0.00
		1-shot	1.61	1.79	1.55	1.12	1.33	1.09	0.00	0.00	0.00	0.00	0.00	0.00
	para	0-shot	6.39	2.04	2.91	4.47	1.44	2.08	0.00	0.00	0.00	0.00	0.00	0.00
		1-shot	6.53	2.16	3.11	2.85	1.23	1.64	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL	sent	sft	12.00	12.34	11.49	9.56	8.95	8.79	0.68	0.49	0.53	1.01	0.82	0.86
	para	sft	4.00	5.84	4.43	6.54	6.01	**5.27**	0.10	0.19	0.11	0.99	0.90	**0.80**
CogVLM2	sent	sft	7.97	7.49	7.22	1.76	1.51	1.50	0.06	0.06	0.06	0.00	0.00	0.00
	para	sft	7.03	3.96	**4.71**	7.49	2.81	3.53	0.00	0.00	0.00	0.00	0.00	0.00
MiniCPM-V	sent	sft	15.97	16.26	15.26	12.74	12.33	11.95	1.08	1.14	**1.00**	1.84	1.79	**1.69**
	para	sft	3.93	5.69	4.31	5.28	5.73	4.83	0.10	0.20	0.13	0.39	0.71	0.48
LLaVA-NeXT	sent	sft	17.06	17.27	**16.23**	14.10	12.88	**12.77**	1.03	0.78	0.84	1.78	1.55	1.54
	para	sft	4.37	5.40	4.55	5.40	5.36	4.81	0.23	0.34	**0.25**	0.59	0.78	0.60

Table 2: Model performance on the DongbaMIE dataset for extracting four semantic element types: Object, Action, Relation, and Attribute. We report Precision (P), Recall (R), and F1-score (F1) as percentages. Evaluations span sentence (Sent) and paragraph (Para) levels. For each element type, yellow highlighting marks the top sentence-level F1-score, while purple highlighting indicates the best paragraph-level F1-score.

flow is shown in Figure 2.

Automated Pre-annotation. We initially employed the DeepSeek v3 API (Liu et al., 2024a) for pre-annotation, extracting four key semantic dimensions—objects, actions, relations, and attributes—from the Chinese translations within the Annotated Collection of Naxi Dongba Manuscripts. This LLM-based step provided a consistent starting point, significantly reducing manual effort. An example extraction prompt is provided in Appendix C.

Manual Review and Refinement. Following pre-annotation, a team of 20 annotators conducted a comprehensive manual review and refinement. This team comprised 19 graduate students in Ethnic Minority Languages and Literature or Computer Science, and one lecturer specializing in Dongba pictograph research. Six members possessed prior experience in Dongba digitalization. All annotators received standardized training on the fundamentals of Dongba pictograms, semantic dimension guidelines, workflow protocols, and the annotation software. All manual tasks were performed using a custom-developed web application (detailed in Appendix A). While standard tools exist, a custom solution was necessitated by the multi-dimensional and highly structured nature of our task. The application's task-oriented workflow ensures semantic integrity when annotators build nested relationships (e.g., relation triplets and attribute pairs) and seamlessly integrates with the LLM pre-annotations for efficient correction. The reviewed results were automatically generated and stored in a flexible JSON format, chosen for its direct compatibility with modern MLLM frameworks. To ensure broader reusability, we will also release the dataset in standard CoNGL and UIMA formats. The focus of this phase was to correct errors and omissions from the LLM, aligning the annotations with expert linguistic understanding.

Multi-level Quality Monitoring. We adopted a three-stage quality control process to ensure consistent and accurate annotations. First, each instance was annotated independently by two annotators in a double-blind setting. Second, a third annotator reviewed and resolved any disagreements through arbitration. Third, two team members continuously monitored the process by randomly reviewing about 5% of the data. When deviations were found, they provided targeted feedback and conducted retraining sessions.

We quantitatively assessed inter-annotator agreement (IAA) using Cohen's Kappa (κ) scores for all four semantic dimensions at the sentence level to evaluate the reliability of the manual annotations. As detailed in Appendix B, the average κ values indicate high consistency, which are 0.803 for the training set, 0.777 for the validation set,

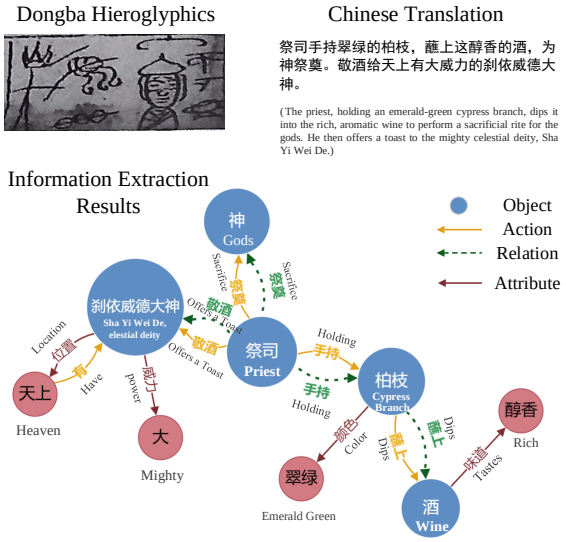


Figure 3: The image presents a semantic visualization yielded by the information extraction framework. Above this visualization, sentence-level Dongba pictograms are shown with their Chinese translations. English descriptions are provided solely for understanding; all annotations are originally in Chinese.

and 0.726 for the test set. The "Action" dimension showed almost perfect agreement ($\kappa > 0.93$), while the "Relation" dimension, reflecting higher semantic complexity, exhibited moderate to substantial agreement. These IAA results confirm the high quality and reliability of our annotation process.

The DongbaMIE dataset is built upon rigorously validated semantic annotations, providing a strong foundation for advancing research in multimodal information extraction for complex, low-resource languages. Detailed statistics are presented in Table 1, and the dataset is released for non-commercial research purposes under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license.

4 Methodology

This section details our experimental methodology for evaluating MLLMs on the DongbaMIE dataset. We conducted zero-shot and one-shot evaluations for leading proprietary MLLMs, specifically GPT-4o (Hurst et al., 2024) and Gemini-2.0 [2], and Supervised Fine-Tuning for a diverse set of open-source MLLMs including LLaVA-NeXT (Liu et al., 2024b), MiniCPM-V-2.6 (Yao et al., 2024), Qwen2-VL (Wang et al., 2024), and CogVLM2 (Hong et al., 2024).

[2] https://deepmind.google/technologies/gemini/

4.1 Zero-shot and Few-shot Evaluation

We evaluated GPT-4o and Gemini-2.0 via Visual Question Answering. In the zero-shot setting, models received only the Dongba image and a predefined prompt targeting the four semantic dimensions of objects, actions, relations, and attributes. For the one-shot setting, a single image-annotation pair was added as an in-context example. As a subset of few-shot, one-shot provides just one example. The specific prompt templates designed for these zero-shot and one-shot settings are visually presented in Appendix D.

4.2 Open-Source MLLMs Supervised Fine-Tuning

We fine-tuned LLaVA-NeXT, MiniCPM-V-2.6, Qwen2-VL, and CogVLM2 on DongbaMIE using instruction-image pairs derived from the dataset. Most models underwent full SFT, where all parameters were updated. For CogVLM2, we employed LoRA, a parameter-efficient fine-tuning method. All models were trained for 3 epochs, with the best-performing checkpoint on a development set selected for testing.

To specifically probe the role of visual representations in this domain, a topic further analyzed in Section 5.3, we applied two distinct SFT strategies to Qwen2-VL. The first was Full SFT, updating all parameters of both its vision and language modules. The second was LLM-only SFT, where the vision encoder was frozen, and only the LLM parameters were fine-tuned. This comparative setup aimed to isolate the contribution of learned, domain-specific visual features to extraction performance.

4.3 Evaluation

Model performance was assessed using Precision (P), Recall (R), and F1-score. For a prediction to be considered correct, we employ a strict Exact Match criterion, meaning the extracted text for an object, action, relation triplet, or attribute pair must perfectly align with the ground-truth annotation. This stringent approach was chosen to rigorously evaluate the model's literal accuracy in interpreting the fine-grained semantics of Dongba pictograms, where precision is vital for linguistic and cultural preservation. These metrics were calculated across both sentence and paragraph levels and for four semantic dimensions: objects, actions, relations, and attributes. Furthermore, we evaluate the model in both single-dimension extraction and concurrent

Model	Level	Mode	Object			Action			Relation			Attribute		
			P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-4o	sent	single	1.73	1.80	1.60	0.34	0.39	0.32	0.00	0.00	0.00	0.00	0.00	0.00
		multi	1.80	1.45	1.46	0.21	0.11	0.14	0.00	0.00	0.00	0.00	0.00	0.00
	para	single	5.91	1.99	2.88	1.07	0.36	0.49	0.00	0.00	0.00	0.00	0.00	0.00
		multi	8.14	1.74	2.74	1.28	0.21	0.33	0.00	0.00	0.00	0.00	0.00	0.00
Gemini-2.0	sent	single	1.85	2.03	1.77	1.04	1.07	0.93	0.00	0.00	0.00	0.00	0.00	0.00
		multi	1.48	1.83	1.50	0.30	0.21	0.23	0.00	0.00	0.00	0.00	0.00	0.00
	para	single	6.39	2.04	2.91	4.47	1.44	2.08	0.00	0.00	0.00	0.00	0.00	0.00
		multi	6.48	1.98	2.91	2.41	0.65	0.99	0.00	0.00	0.00	0.00	0.00	0.00
Qwen2-VL	sent	single	12.00	12.34	11.49	9.56	8.95	8.79	0.68	0.49	0.53	1.01	0.82	0.86
		multi	11.80	12.11	11.31	6.20	5.85	5.63	0.66	0.61	0.57	1.48	1.33	1.31
	para	single	4.00	5.84	4.43	6.54	6.01	5.27	0.10	0.19	0.11	0.99	0.90	0.80
		multi	4.65	6.73	5.14	5.18	6.84	5.14	0.03	0.05	0.04	0.16	0.27	0.19
CogVLM2	sent	single	7.97	7.49	7.22	1.76	1.51	1.50	0.06	0.06	0.06	0.00	0.00	0.00
		multi	7.75	8.63	7.65	2.70	2.52	2.40	0.67	0.62	0.58	0.13	0.17	0.14
	para	single	7.03	3.96	4.71	7.49	2.81	3.53	0.00	0.00	0.00	0.00	0.00	0.00
		multi	7.94	3.85	4.74	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MiniCPM-V	sent	single	15.97	16.26	15.26	12.74	12.33	11.95	1.08	1.14	1.00	1.84	1.79	1.69
		multi	4.40	5.06	4.41	2.44	2.25	2.11	0.19	0.21	0.18	0.44	0.30	0.32
	para	single	3.93	5.69	4.31	5.28	5.73	4.83	0.10	0.20	0.13	0.39	0.71	0.48
		multi	3.94	5.82	4.40	5.41	6.61	5.22	0.05	0.11	0.07	0.13	0.19	0.13
LLaVA-NeXT	sent	single	17.06	17.27	16.23	14.10	12.88	12.77	1.03	0.78	0.84	1.78	1.55	1.54
		multi	16.00	16.44	15.32	9.30	9.15	8.76	1.12	1.19	1.04	1.71	1.79	1.62
	para	single	4.37	5.40	4.55	5.40	5.36	4.81	0.23	0.34	0.25	0.59	0.78	0.60
		multi	4.05	5.27	4.28	3.72	3.71	3.18	0.12	0.20	0.12	0.29	0.39	0.30

Table 3: Comparison of the performance of MLLMs in extracting single semantic dimensions and multiple concurrent semantic dimensions from Dongba hieroglyphs. Results are analyzed at the sentence and paragraph level, where P, R and F1 denote precision, recall and F1 score, respectively.

multi-semantic extraction to provide a comprehensive benchmark.

5 Experimental Analysis

5.1 Overall Performance

Extracting multimodal semantic information from the DongbaMIE dataset is challenging for current MLLMs, as shown in Table 2. Initial evaluations revealed significant limitations in proprietary models like GPT-4o and Gemini-2.0 using zero-shot and one-shot settings. For instance, GPT-4o's 0-shot sentence-level object extraction F1 was only 1.60. Both models were largely ineffective for relation and attribute extraction, achieving F1 scores of 0.00. When given paragraph-level input or a 1-shot prompt, these models achieved only minor F1 score increases for object and action extraction. Such improvements were far from overcoming their core performance deficiencies, indicating their struggle to accurately interpret the detailed semantics of Dongba pictograms.

SFT on open-source MLLMs led to varied degrees of performance improvement. At the sentence level, LLaVA-NeXT performed best, achieving top F1 scores of 16.23 for Object and 12.77 for Action extraction. MiniCPM-V-2.6 was also competitive, excelling with a sentence-level Attribute F1 score of 1.69 and achieving strong Object and Action F1 scores. While Qwen2-VL improved over zero-shot results, LLaVA-NeXT and MiniCPM-V-2.6 surpassed it. CogVLM2, using parameter-efficient fine-tuning, underperformed significantly, especially for relation and attribute extraction. This may indicate the necessity of full fine-tuning for this domain.

Across all evaluated models, two prominent findings emerge. First, sentence-level extraction consistently surpasses paragraph-level performance, indicating that MLLMs struggle with longer pictographic contexts. This performance decline is attributable to compounded challenges in longer sequences. The difficulty in recognizing individual pictographs leads to errors that propagate and accumulate, fracturing the model's grasp of long-range visual-semantic dependencies. Concurrently, paragraphs present increased narrative complexity and visual density, which MLLMs struggle to parse accurately without the guideposts of a fixed grammatical structure. Second, relation and attribute extraction remain exceptionally challenging.

Figure 4: The image displays a Dongba pictograph with manually added annotations highlighting Objects (blue) and Actions (yellow). The text below shows multimodal semantic extraction results from six models, alongside ground truth labels. Prediction and Ground truth are shown in Chinese, with English notes for clarity.

Even top-performing models like LLaVA-NeXT and MiniCPM-V-2.6 achieved F1 scores of 1.69 or less for these tasks. This highlights a fundamental limitation: MLLMs struggle with the structured semantics inherent in these tasks. Future work can address this by integrating graph-based reasoning modules such as GNNs (Zhou et al., 2020; Han et al., 2025) to model structural dependencies, alongside enhanced visual-semantic encoders and the infusion of domain-specific knowledge graphs.

5.2 Single vs. Multi-Semantic Extraction

This section evaluates the multi-extraction performance of MLLMs in four semantic dimensions. We compare this with the single dimension extraction performance. Table 3 details the results.

Zero-shot models lack task-specific training. They typically struggle with concurrent multi-semantic extraction. For instance, GPT-4o's F1 scores generally decreased across sentence and paragraph levels. Its sentence-level Object F1 score dropped from 1.60 to 1.46. Gemini's performance also markedly declined, especially in Action extraction. Its sentence-level F1 score plummeted from 0.93 to 0.23. For both models, the already challenging relation and attribute extraction showed no improvement in multi-task settings. Their F1 scores remained zero. This underscores the inherent difficulty of concurrent complex semantic extraction without targeted training.

Models undergoing SFT showed varied performance in concurrent multi-semantic extraction. This task remained challenging for them. For CogVLM2, at the sentence level, the Object F1 score improved from 7.22 to 7.65, and the Action F1 score rose from 1.50 to 2.40. Previously unrecognized relation information was also extracted, achieving an F1 score of 0.58, along with attribute information, which reached an F1 score of 0.14. For performance at the paragraph level, there were varying degrees of increases or decreases. Other SFT models also displayed such inconsistencies. Improvements in some aspects or contexts were often offset by declines in others. Thus, SFT provides situational benefits but does not fully resolve concurrent extraction issues, frequently leading to improvements in some areas alongside performance drops in others.

5.3 Impact of Visual Feature Learning

We evaluated the impact of visual representation learning on model performance. To do this, we compared two Qwen2-VL setups: one with only the LLM fine-tuned, and another with full fine-tuning, including its visual module. Table 4 presents these results. The findings clearly demonstrate that full fine-tuning significantly enhances performance in Dongba pictograph multimodal semantic extraction. This improvement is particularly evident at the sentence level. For instance, the F1 score for Object extraction substantially increased from 11.49 to 18.06. Similarly, the Action extraction F1 score rose from 8.79 to 15.58. These results directly show that the model's ability to learn and optimize visual features specific to Dongba pictograms is crucial for accurately understanding its pictographic semantics. Conversely, limited visual feature extraction capabilities are a key factor contributing to performance bottlenecks.

Full fine-tuning also yielded performance gains at the paragraph level. For example, Object F1 improved from 4.43 to 6.43. However, the extent of this improvement was less pronounced than at the sentence level. This observation may suggest that the contribution of high-quality visual features diminishes in longer textual contexts. Alternatively, current model architectures might still face limitations in effectively modeling long-range visual dependencies. Overall, these findings underscore the critical importance of targeted enhancements in visual feature learning for processing pictographic

Model	Level	Object			Action			Relation			Attribute		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Qwen2-VL (llm-sft)	sent	12.00	12.34	11.49	9.56	8.95	8.79	0.68	0.49	0.53	1.01	0.82	0.86
	para	4.00	5.84	4.43	6.54	6.01	5.27	0.10	0.19	0.11	0.99	0.90	0.80
Qwen2-VL (full-sft)	sent	18.91	18.94	18.06	17.26	15.47	15.58	1.04	0.73	0.76	2.18	1.81	1.86
	para	6.47	7.12	6.43	6.71	6.38	5.76	0.12	0.05	0.07	0.65	0.77	0.63

Table 4: Fine-tuning of Qwen2-VL for semantic information extraction from Dongba hieroglyphs: comparison of LLM (llm-sft) model fine-tuning only vs. full model fine-tuning (full-sft, including visual module) in MLLMs. We report Precision (P), Recall (R), and F1-score (F1) as percentages.

scripts like Dongba. They also provide clear directions for future work aimed at improving model capabilities.

5.4 Error Analysis and Future Directions

This section presents a qualitative case study based on the Dongba pictograph in Figure 4. It highlights key challenges faced by MLLMs in interpreting Dongba hieroglyphs. The analysis focuses on extracting objects and actions, which are essential for narrative understanding in hieroglyphic scripts.

Many models, including Gemini 2.0 and CogVLM2, failed to recognize core environmental objects such as "mountain" and "water." Others, like Qwen2-VL and LLaVA-NeXT, replaced generic symbols with overly specific named entities, such as "Le Qi He Mu mountain." MiniCPM-V showed better performance in extracting key objects.

Interpreting symbolic features proved problematic. This included erroneous segmentation of unified entities (GPT-4o) and overly abstract readings of potential human figures as "village god" (CogVLM2, MiniCPM-V).

Action extraction was acutely challenging. Models predominantly returned inaccurate, abstract labels (Gemini 2.0: "sacrifice"; GPT-4o: "worship"; MiniCPM-V: "welcome"; LLaVA-NeXT: "remove") or no output (Qwen2-VL, CogVLM2), failing to identify specific activities such as "gathering firewood."

A core issue underlying these varied errors is the model's struggle with semantic granularity, particularly in managing hypernym-hyponym relationships. This is not merely a recognition failure but an inability to navigate the abstract-to-concrete hierarchy inherent in Dongba's pictorial ideographs, where a symbol's precise meaning is determined by context. This indicates that MLLMs lack a deep understanding of the Dongba pictograms as a unique knowledge system. Addressing this will require future work on multi-granularity semantic learning, integrating external cultural knowledge, and developing context-aware inference mechanisms tailored to ideographic features.

6 Conclusion

In this study, we introduced DongbaMIE, the first multimodal information extraction dataset for Dongba pictograms. It aims to advance the processing of endangered scripts. By evaluating against state-of-the-art closed-source and open-source MLLMs, we found the limitations of zero-shot and few-shot methods, while demonstrating the potential of supervised fine-tuning. Despite this, current models still struggle to understand the complexity of Dongba pictograms. In the future, we will focus on expanding the dataset annotations with more fine-grained annotations. At the same time, we will improve the multimodal representation to further enhance the ability to extract information from Dongba pictograms. Our dataset will be released soon after the acceptance of the paper. This is intended to facilitate reproducible research in the field of endangered text conservation.

Limitations

Despite the innovative multimodal dataset presented in this study, there are still several limitations that need to be addressed. First, due to the limitations of current multimodal models in handling low-resource languages and pictographic scripts, these models still face significant challenges in accurately extracting semantic information from Dongba pictograms. In particular, the model performance remains limited when interpreting the actions, relations, and attributes conveyed by Dongba symbols. We believe that with the continued advancement of MLLMs, future models will be better equipped to handle complex image and semantic information, thereby improving the semantic un-

derstanding and information extraction of Dongba pictograms. Second, although the dataset covers multiple image levels and four semantic dimensions, we plan to expand the dataset in future work to explore more fine-grained symbol and semantic annotations, such as context-based semantic understanding of individual Dongba pictograph characters. Additionally, integrating data from other modalities, such as audio, video, or depth images, could provide the model with more contextual information, thus enhancing overall multimodal understanding capabilities.

Acknowledgements

We would like to thank the following individuals for their participation in the review of the DongbaMIE dataset (in no particular order): Weizheng Qiao, Ziwei Sun, Wenhao Tao, Di Du, Yongxu Tao, Jun Jiang, Hao Li, Yuzhan Li, Dan Wu, Xiaoyuan Ma, Moxuan Xu, Zhongyi Fan, Fayi Li, Jiaqi Chen, Chengyang Bao, Shuo Wang, Jingliang Liu, Quanyi Ou, Yizhi Ma and Jiayue Wang.

This work was supported by National Natural Science Foundation of China (Grant No. 62236011), National Social Science Foundation of China (Grant No. 20&ZD279).

References

Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, and Marco C. Passarotti, editors. 2023. *Proceedings of the Ancient Language Processing Workshop*. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria.

Xiaojun Bi and Yanlong Luo. 2024. Incomplete handwritten dongba character image recognition by multiscale feature restoration. *Heritage Science*, 12(1):218.

Rina Buoy, Masakazu Iwamura, Sovila Srun, and Koichi Kise. 2023. Toward a low-resource non-latin-complete baseline: An exploration of khmer optical character recognition. *IEEE Access*, 11:128044–128060.

Jacob Carlson, Tom Bryan, and Melissa Dell. 2024. Efficient OCR for building a diverse digital history. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8105–8115, Bangkok, Thailand. Association for Computational Linguistics.

Danlu Chen, Freda Shi, Aditi Agarwal, Jacobo Myerston, and Taylor Berg-Kirkpatrick. 2024. LogogramNLP: Comparing visual and textual representations of ancient logographic writing systems for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14238–14254, Bangkok, Thailand. Association for Computational Linguistics.

Zijian Chen, Tingzhu Chen, Wenjun Zhang, and Guangtao Zhai. 2025. Obi-bench: Can lmms aid in study of ancient script on oracle bones? *Preprint*, arXiv:2412.01175.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Shruti Daggumati and Peter Z. Revesz. 2018. Data mining ancient script image data using convolutional neural networks. In *Proceedings of the 22nd International Database Engineering & Applications Symposium*, IDEAS '18, page 267–272, New York, NY, USA. Association for Computing Machinery.

Mattia De Cao, Nicola De Cao, Angelo Colonna, and Alessandro Lenci. 2024. Deep learning meets egyptology: a hieroglyphic transformer for translating Ancient Egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 71–86, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Li Deng and Yang Liu. 2018. *Deep learning in natural language processing*. Springer.

Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi, and Jiang Gui. 2025. Temporal working memory: Query-guided segment refinement for enhanced multimodal understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3393–3409, Albuquerque, New Mexico. Association for Computational Linguistics.

Hassan El Bahi. 2024. Handwritten text recognition and information extraction from ancient manuscripts using deep convolutional and recurrent neural network. *Soft Computing*, 28(20):12249–12268.

Shengxiang Gao, Zhiwen Tang, Zhengtao Yu, Chao Liu, and Lin Wu. 2018. Chinese-naxi syntactic statistical machine translation based on tree-to-tree. *International Journal of Information and Communication Technology*, 13(3):351–360.

Shengxiang Gao, Xiuzhen Yang, Zhengtao Yu, Xiao Pan, and Jianyi Guo. 2017. Chinese-naxi machine translation method based on naxi dependency language model. *International Journal of Machine Learning and Cybernetics*, 8:333–342.

Shai Gordin, Morris Alper, Avital Romach, Luis Saenz Santos, Naama Yochai, and Roey Lalazar. 2024. CuReD: Deep learning optical character recognition for cuneiform text editions and legacy materials. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 130–140, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Sanjana Gunna, Rohit Saluja, and C. V. Jawahar. 2021. Towards boosting the accuracy of non-latin scene text recognition. In *Document Analysis and Recognition – ICDAR 2021 Workshops*, pages 282–293, Cham. Springer International Publishing.

Jiaqi Han, Jiacheng Cen, Liming Wu, Zongzhao Li, Xiangzhe Kong, Rui Jiao, Ziyang Yu, Tingyang Xu, Fandi Wu, Zihe Wang, et al. 2025. A survey of geometric graph neural networks: Data structures, models and applications. *Frontiers of Computer Science*, 19(11):1911375.

Wanbao He and Jiaxiu He. 1999. *An Annotated Collection of Naxi Dongba Manuscripts*. Yunnan People's Publishing House, Kunming.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Yuxiao Dong, and Jie Tang. 2024. Cogvlm2: Visual language models for image and video understanding. *Preprint*, arXiv:2408.16500.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022. OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

Kai Jin, Dan Zhao, and Wuying Liu. 2023. Morphological and semantic evaluation of Ancient Chinese machine translation. In *Proceedings of the Ancient Language Processing Workshop*, pages 96–102, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kyeongpil Kang, Kyohoon Jin, Soyoung Yang, Soojin Jang, Jaegul Choo, and Youngbin Kim. 2021. Restoring and mining the records of the Joseon dynasty via neural language modeling and machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4031–4042, Online. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 32–39, Minneapolis, Minnesota. Association for Computational Linguistics.

Yaohan Lu. 2024. Application of ai in the field of documentary heritage: A review of the literature. *Journal of Artificial Intelligence Research*, 1(2):15–21.

Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. 2023a. Geolayoutlm: Geometric pre-training for visual information extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7092–7101.

Jiaming Luo, Frederik Hartmann, Enrico Santus, Regina Barzilay, and Yuan Cao. 2021. Deciphering undersegmented ancient scripts using phonetic prior. *Transactions of the Association for Computational Linguistics*, 9:69–81.

Yanlong Luo, Yiwen Sun, and Xiaojun Bi. 2023b. Multiple attentional aggregation network for handwritten dongba character recognition. *Expert Systems with Applications*, 213:118865.

Yuqi Ma, Shanxiong Chen, Yongbo Li, Jingliu He, Qiuyue Ruan, Wenjun Xiao, Hailing Xiong, and XiaoLiang Li. 2024. Stef: a swin transformer-based enhanced feature pyramid fusion model for dongba character detection. *Heritage Science*, 12(1):206.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

John Pavlopoulos, Thea Sommerschield, Yannis Assael, Shai Gordin, Kyunghyun Cho, Marco Passarotti, Rachele Sprugnoli, Yudong Liu, Bin Li, and Adam Anderson, editors. 2024. *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*. Association for Computational Linguistics, Hybrid in Bangkok, Thailand and online.

John Pavlopoulos, Alexandros Xenos, and Davide Picca. 2022. Sentiment analysis of Homeric text: The 1st book of Iliad. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7071–7077, Marseille, France. European Language Resources Association.

Aleksi Sahala, Miikka Silfverberg, Antti Arppe, and Krister Lindén. 2020. Automated phonological transcription of Akkadian cuneiform text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3528–3534, Marseille, France. European Language Resources Association.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, pages 703–747.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the EvaLatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia. ELRA and ICCL.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Lin Sun, Kai Zhang, Qingyuan Li, and Renze Lou. 2024. Umie: Unified multimodal information extraction with instruction tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19062–19070.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1:5–21.

JM Wang, Sh B Wang, and SP Cheng. 2011. Research on the digital protection strategies of intangible cultural heritage. *Softw. Guide*, 10(8):49–51.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Duoduo Xu. 2023. Digital approaches to understanding dongba pictographs. *International Journal of Digital humanities*, 4(1):131–146.

Soyoung Yang, Minseok Choi, Youngwoo Cho, and Jaegul Choo. 2023. HistRED: A historical document-level relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3207–3224, Toronto, Canada. Association for Computational Linguistics.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Haneul Yoo, Jiho Jin, Juhee Son, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. HUE: Pretrained model and dataset for understanding hanja documents of Ancient Korea. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1832–1844, Seattle, United States. Association for Computational Linguistics.

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic translation alignment for Ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.

A Data Annotation and Review Web Application

We employed a custom-developed web application (Figure 8) to manually review and refine pre-annotated Dongba pictograph data. This application standardizes the annotation workflow and improves data quality. It provides a visual interface for Dongba pictograms and their multimodal information. The application enables straightforward editing of pre-annotations to correct machine-generated errors and omissions. It also guides users according to established annotation protocols and automatically stores reviewed results in JSON format on the server-side.

B Inter-Annotator Agreement Scores

To ensure the quality and reliability of our manual annotations, we conducted a quantitative evaluation of Inter-Annotator Agreement (IAA) at the sentence level. This evaluation covered all four core semantic dimensions: Action, Object, Relation, and Attribute. We employed Cohen's Kappa (κ) coefficient (Cohen, 1960) as the evaluation metric. Cohen's Kappa is widely used to measure agreement between annotators, accounting for agreement that could occur by chance (McHugh, 2012). The IAA results are presented in Table 5.

Overall Agreement: The average κ values are 0.803 for the training set, 0.777 for the development set, and 0.726 for the test set. According to the standard interpretation of Kappa values (Landis and Koch, 1977), these scores indicate substantial to almost perfect agreement. This suggests a high overall reliability for our manual annotations.

Dimensional Differences: The Action dimension exhibits very high agreement across all datasets. Its κ values consistently exceed 0.93 (Train: 0.981, Dev: 0.976, Test: 0.935), signifying almost perfect agreement. This indicates a strong consensus among annotators in identifying actions. In contrast, the Relation dimension shows lower κ values (Train: 0.671, Dev: 0.626, Test: 0.542). These scores fall within the moderate to substantial agreement range. This lower agreement likely reflects the inherent semantic complexity and subjectivity of the Relation dimension. Annotators may face more challenges in interpreting and labeling complex relationships, leading to more divergence. Nevertheless, these values still indicate an acceptable quality of annotation for this dimension. The Object and Attribute dimensions also demonstrate good agreement. For instance, in the training set, the κ values for Object (0.734) and Attribute (0.825) suggest consistent judgments by annotators.

Cross-Dataset Comparison: The average IAA for the training set (0.803) is slightly higher than those for the dev (0.777) and test (0.726) sets. These minor variations are normal. They may reflect subtle differences in data distribution or slight fluctuations in annotator performance across batches. However, the overall trend indicates that a high level of annotation quality was maintained across all datasets.

DeepSeek v3 IE Prompt

You are a semantic analysis assistant. Your task is to analyze an input Dongba pictograph corpus and extract values corresponding to four predefined semantic labels. The output must be a structured JSON object. The four semantic labels are defined as follows:

1. Action: List all involved actions, such as behaviors or dynamic events. This should not include information about the objects participating in the actions.
2. Object: List all minimal entity units. This should not include their attribute information.
3. Relation: Describe semantic associations between objects, such as positional, possessive, or functional relationships. Represent these using a triplet structure: (entity1, relation, entity2).
4. Attribute: Extract specific characteristics or descriptions of objects, such as color, function, or state.

Output Requirements:

- All generated content must be based directly on the input corpus. Do not add extraneous information.
- If a value for a label cannot be determined, or if the input corpus lacks content for a specific label, return an empty array ([]). Do not make arbitrary inferences.
- Strictly follow the prescribed order of labels in the JSON output. Ensure consistency and adhere to all formatting specifications.

JSON Format Example:

```
{
  "Action": [ {"action": "..."}],
  "Object": [{"object": "..."}],
  "Relation": [{"entity1": "...", "relation": "...", "entity2": "..."}],
  "Attribute": [{ "object": "...", "attribute": "...", "value": "..."}]
}
```

The input Dongba pictograph corpus is as follows:

Figure 5: DeepSeek v3 IE Prompt template automates multi-label semantic analysis of Dongba pictograms, outputting actions, objects, relations, and attributes in structured JSON.

C Example Prompt for Automated Pre-annotation

During automated pre-annotation, we utilized the DeepSeek v3 LLM API to extract key semantic dimensions from Chinese translations in the Naxi Dongba manuscript annotation corpus. Figure 5 shows an example prompt designed to guide the model in this information extraction task. This prompt directs the model to identify and extract predefined semantic elements: entities, actions, relations, and attributes.

Specific guidelines within the prompt addressed nuanced aspects, such as defining the scope of actions or the contextual nature of entity attributes. To streamline downstream processing and support subsequent review by human annotators, the prompt requested output in a structured format (e.g., JSON).

Split	Sentence	Action	Object	Relation	Attribute	Avg.
Train	18824	0.981	0.734	0.671	0.825	0.803
Dev	2,353	0.976	0.751	0.626	0.753	0.777
Test	2,353	0.935	0.693	0.542	0.733	0.726

Table 5: Cohen's Kappa (κ) Inter-Annotator Agreement (IAA) scores for each dataset (Train, Dev, Test), covering sentence counts, scores for each semantic dimension (Action, Object, Relation, Attribute), and average (Avg.) κ values.

Zero-shot Prompt

You are an expert system specializing in Dongba pictographs. Your task is to extract structured information from provided Dongba pictographs, focusing on the aspect detailed below.

Information to be Extracted:

1. Object: Identify the primary entities depicted in the image.

Constraints and Guidelines:

1. Dongba script consists of highly abstract pictographs. Semantic interpretation requires inferring meaning from the morphological features of the graphic symbols.
2. The output must be strictly in JSON format, without any additional explanations or natural language descriptions.
3. The key name must strictly conform to the specified convention (i.e., "Object").
4. The value associated with the "Object" key must be an array, even if it contains only a single element or no elements.
5. If no objects are identified, return an empty array ([]) as the value for the "Object" key.
6. The textual names of extracted objects (the values within the JSON structure) must be in Chinese.
7. All extracted object names listed within the array must be unique; duplicates are not permitted.
8. The JSON output must use double quotes for all keys and string values, as per standard JSON formatting.

Output JSON Format:

"Object": [{"object": "..."}]

The input Dongba pictographs is as follows:

Figure 6: Zero-shot prompt template for MLLMs (e.g., GPT-4o, Gemini) to extract four semantic types from Dongba hieroglyphs.

Few-shot Prompt

You are an expert system specializing in Dongba pictographs. Your task is to extract structured information from provided Dongba pictographs, focusing on the aspect detailed below.

Information to be Extracted:

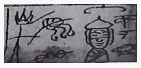
1. Object: Identify the primary entities depicted in the image.

Constraints and Guidelines:

1. Dongba script consists of highly abstract pictographs. Semantic interpretation requires inferring meaning from the morphological features of the graphic symbols.
2. The output must be strictly in JSON format, without any additional explanations or natural language descriptions.
3. The key name must strictly conform to the specified convention (i.e., "Object").
4.The value associated with the "Object" key must be an array, even if it contains only a single element or no elements.
5.If no objects are identified, return an empty array ([]) as the value for the "Object" key.
6.The textual names of extracted objects (the values within the JSON structure) must be in Chinese.
7. All extracted object names listed within the array must be unique; duplicates are not permitted.
8. The JSON output must use double quotes for all keys and string values, as per standard JSON formatting.

Output JSON Format:

"Object": [{"object": "..."}]



Example (note: the example values provided are in Chinese, as per constraint #6):

"Object": [{"object": "祭司"}, {"object": "柏枝"}, {"object": "酒"}, {"object": "神"}, {"object": "刹依威德大神"}]

The input Dongba pictographs is as follows:

Figure 7: Few-shot prompt template for MLLMs (e.g., GPT-4o, Gemini) to extract four semantic types from Dongba pictograms.

The prompt was carefully constructed with the goal of maximizing the accuracy and completeness of the semantic information captured from the translations.

D Prompt Templates

This appendix presents prompt templates used to evaluate two proprietary MLLMs—GPT-4o and Gemini 2.0—on Dongba pictograms information extraction (Section 4.1). These templates are presented in English for clarity. The operational versions used in our evaluations were authored and deployed in Chinese.

Figure 6 shows the template for zero-shot semantic information extraction from Dongba pictographic. Applied to MLLMs like GPT-4o and Gemini 2.0, this prompt directs the model to identify and structure multiple semantic categories (e.g., actions, objects, relations, attributes). This process relies on instructions alone, without in-context examples.

Figure 7 illustrates the "Few-shot Prompt" template for one-shot (an instance of few-shot learning) diverse semantic extraction from Dongba pictographic images. This template includes an in-prompt example with a placeholder for an image-annotation pair. This guides the model to identify semantic elements (e.g., actions, objects, relations, attributes) from visual input.

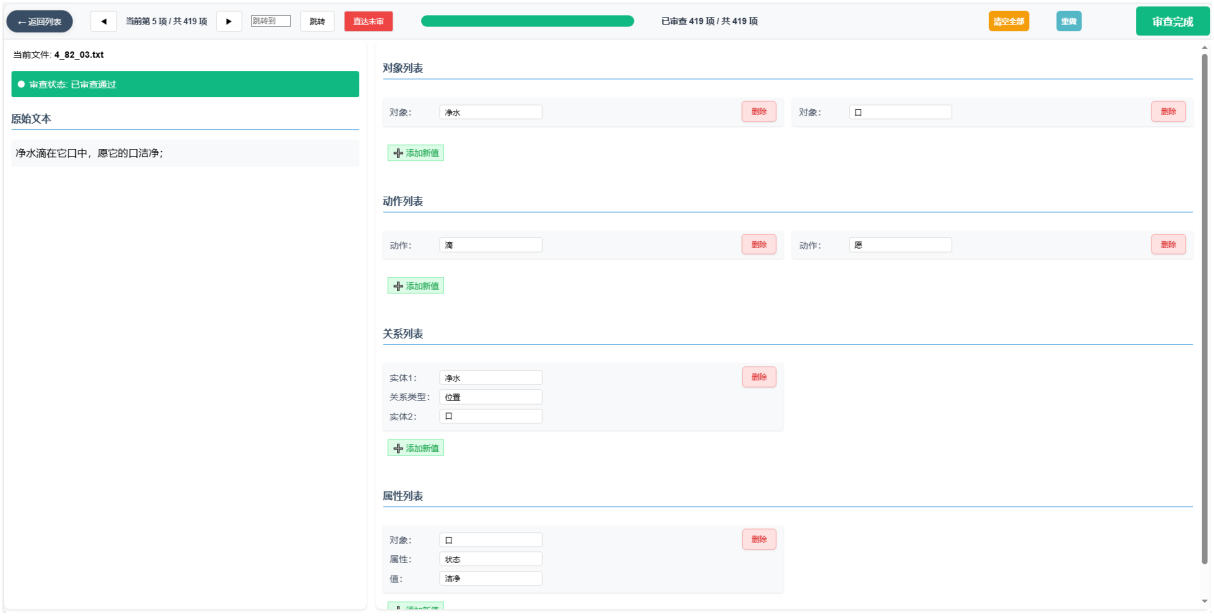


Figure 8: This is a web application page for manual review of Dongba pictographic information extraction.