



Time Travel: A Comprehensive Benchmark to Evaluate LMMs on Historical and Cultural Artifacts

Sara Ghaboura^{1†} Ketan More^{1†} Ritesh Thawkar¹ Wafa Alghallabi¹ Omkar Thawakar¹
 Fahad Shahbaz Khan^{1,2} Hisham Cholakkal¹ Salman Khan^{1,3} Rao Muhammad Anwer^{1,4}

¹Mohamed bin Zayed University of AI, ²Linköping University, ³Australian National University, ⁴Aalto University
 {sara.ghaboura, ketan.more, omkar.thawakar}@mbzuai.ac.ae

<https://mbzuai-oryx.github.io/TimeTravel/>

Abstract

Understanding historical and cultural artifacts demands human expertise and advanced computational techniques, yet the process remains complex and time-intensive. While large multimodal models offer promising support, their evaluation and improvement require a standardized benchmark. To address this, we introduce *TimeTravel*, a benchmark of 10,250 expert-verified samples spanning 266 distinct cultures across 10 major historical regions. Designed for AI-driven analysis of manuscripts, artworks, inscriptions, and archaeological discoveries, *TimeTravel* provides a structured dataset and robust evaluation framework to assess AI models' capabilities in classification, interpretation, and historical comprehension. By integrating AI with historical research, *TimeTravel* fosters AI-powered tools for historians, archaeologists, researchers, and cultural tourists to extract valuable insights while ensuring technology contributes meaningfully to historical discovery and cultural heritage preservation. We evaluate contemporary AI models on *TimeTravel*, highlighting their strengths and identifying areas for improvement. Our goal is to establish AI as a reliable partner in preserving cultural heritage, ensuring that technological advancements contribute meaningfully to historical discovery. Our code is available at: <https://github.com/mbzuai-oryx/TimeTravel>.

1 Introduction

In recent years, Large Multimodal Models (LMMs) have made significant strides in visual reasoning, perception, and multimodal understanding. Models such as GPT-4V (OpenAI, 2024) and LLaVA (Liu et al., 2023) have excelled in image caption-

[†]Equal contribution.

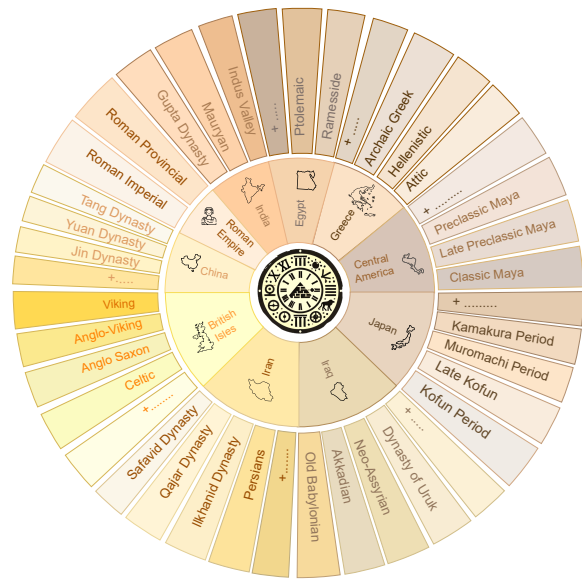


Figure 1: *TimeTravel* Taxonomy categorizes artifacts from 10 major civilizations, spanning diverse historical and prehistoric periods. It encompasses 266 distinct cultures and over 10k manually verified historical artifact samples, providing a structured framework for comprehensive AI-driven analysis.

ing, visual question answering (VQA), and complex visual reasoning, driving the development of benchmarks (Chiu et al., 2024; Nayak et al., 2024; Alwajih et al., 2024) to assess their capabilities. These benchmarks predominantly focus on modern objects, cultural landmarks, and textual sources, extending multimodal AI applications to domains such as medical imaging, remote sensing, and real-world scene understanding (Ghaboura et al., 2025). However, a critical gap remains—LMMs fail to address the historical dimension of visual data, particularly artifacts that shaped human civilization.

Historical artifacts, from ancient manuscripts and inscriptions to architectural ruins and cultural symbols, offer invaluable insights into the evolution



Figure 2: **TimeTravel Samples**. Showcasing diverse cultural representations from various regions across the globe, these examples span multiple artifact categories, including coins, accessories, tools, and statues from ancient civilizations. Each artifact is accompanied by a detailed description, providing valuable contextual and historical insights. Additional TimeTravel examples can be found in Fig.7 and Fig.8.

of societies, artistic expression, and technological advancements. These artifacts preserve cultural heritage and serve as primary sources for understanding belief systems, trade networks, and socio-political structures of past civilizations. However, interpreting them requires deep contextual knowledge, which current LMMs struggle to achieve, particularly in non-English and non-Western historical contexts. While some models have been extended to low-resource languages to bridge cultural gaps (Heakl et al., 2025), they lack systematic capabilities to analyze artifacts from diverse civilizations. This limitation highlights the urgent need for a specialized benchmark that evaluates AI’s ability to process and understand historical artifacts with cultural and temporal awareness.

To address this challenge, we introduce TimeTravel, an open-source comprehensive benchmark (see Table 1) for evaluating LMM performance in historical artifact analysis across diverse civilizations. TimeTravel encompasses several major ancient and prehistoric civilizations across 10 distinct regions, spanning 266 cultural groups. It offers a structured taxonomy tailored for AI-driven historical research (see Fig. 1). Unlike existing benchmarks that focus on generic object recognition, TimeTravel prioritizes historical knowledge, contextual reasoning, and cultural preservation, making it a pioneering effort in multimodal AI evaluation. The benchmark consists of over 10k curated samples, each accompanied by high-quality images of manuscripts, inscriptions, sculptures, paintings, and archaeological discoveries. These samples assess key aspects of multimodal understanding, including visual perception, contextual reasoning, and cross-civilizational knowledge. Meticulously verified by historians and archaeologists, the dataset ensures accuracy, cultural relevance, and historical integrity. By evaluating both closed- and open-source LMMs

on TimeTravel, we aim to identify their strengths and limitations in handling historically significant artifacts, paving the way for AI models that contribute meaningfully to cultural heritage preservation and historical analysis.

Domain	British Museum	MMMU	Oracle-MNIST	Ithaca	Kao Kore	HUST-OBS	TimeTravel (ours)
Hist. Artifact Recog.	✓	✗	✗	✗	✓	✗	✓
Geographic Region	✓	✗	✗	✓	✓	✗	✓
Ancient Artifacts	✓	✗	✗	✗	✗	✗	✓
Contextual History	✗	✗	✗	✗	✗	✗	✓
Image-Text Pairs	✓	✓	✗	✗	✗	✗	✓
Open-Source	✗	✓	✓	✗	✓	✓	✓

Table 1: The comparison of datasets and benchmarks for historical and cultural artifacts, evaluating features like **artifact recognition**, **geographic coverage**, **multimodal understanding**, and **metadata inclusion** with existing data such as British Museum (Tully, 2020), MMMU (Yue et al., 2024), Oracle-MNIST (Wang and Deng, 2022), Ithaca (Assael et al., 2022), KaoKore (Tian et al., 2020), HUST-OBS (Wang et al., 2024). TimeTravel stands out as the most comprehensive benchmark, uniquely integrating multimodal data, historical context, and a dedicated focus on ancient artifacts to support AI-driven cultural heritage research.

2 The TimeTravel Dataset

2.1 Data Collection

Our research is based on a well-structured and meticulously curated dataset sourced from museum collections, which houses an extensive collection of artifacts from diverse civilizations. From this vast repository, we compiled a dataset spanning 266 cultural groups, allowing the analysis of cultural, technological, and social developments over a broad historical timeline.

To ensure the integrity of our benchmark, we followed a systematic data collection process. We first identified key civilizations and historical periods relevant to our study, then collaborated closely with experts to validate the authenticity and completeness of each record. As a result, our dataset com-

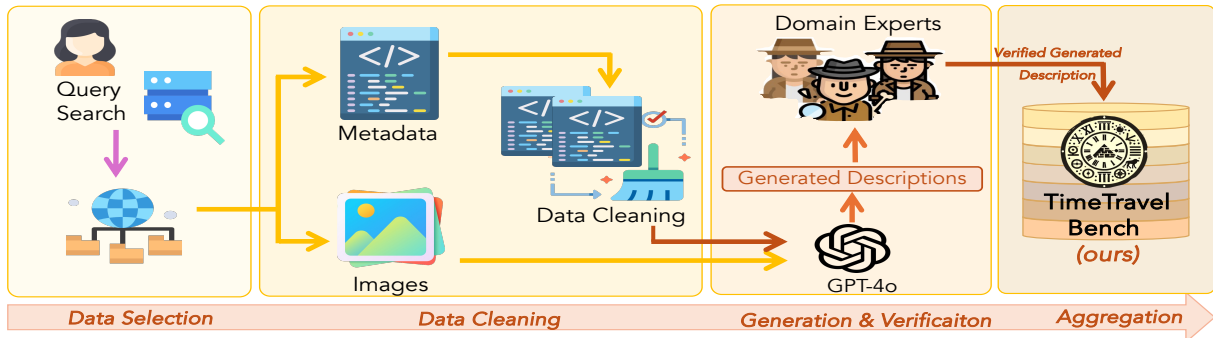


Figure 3: **TimeTravel Data Pipeline.** A structured workflow that collects image and text data from museum websites, cleans metadata, and integrates it with visual content. The GPT-4o model generates detailed, context-aware descriptions, which are refined by experts for accuracy before forming the TimeTravel Benchmark.

prises 10,250 carefully curated samples (see Fig 2). Each entry—ranging from artifacts and inscriptions to ancient manuscripts—was meticulously verified by historians and archaeologists, ensuring accuracy and reliability. By incorporating data from multiple civilizations, our benchmark provides a diverse and comprehensive perspective, avoiding the limitations of a single historical narrative while preserving the historical context for in-depth analysis. This meticulous approach allows us to reveal significant patterns in human history, offering valuable insights into the evolution of human history and civilizations over time.

2.2 Image-Text pair Generation

The dataset features a diverse range of historical objects, ensuring comprehensive documentation and contextual understanding. However, many metadata fields—such as title, iconography, and date—were missing or incomplete. To address this, we employed GPT-4o to generate detailed, context-aware textual descriptions based on the available metadata (see Fig. 5 and 6). To further enhance usability, we structured these descriptions into image-text pairs, ensuring that each artifact is not only visually documented but also enriched with contextual and cultural insights. By improving multimodal model compatibility and supporting digital archiving, this approach strengthens research in cultural heritage preservation while bridging gaps in existing records.

2.3 Data Filtering and Verification

To guarantee the accuracy and reliability of our dataset, we implemented a rigorous data filtering and verification process (Fig. 3). This process combined manual expert validation with automated techniques to eliminate inconsistencies, fill in miss-

ing details where possible, and authenticate historical records. During data cleaning, we addressed missing or incomplete metadata—such as titles, dates, and iconography—by cross-referencing museum archives, academic sources, and expert insights. Unavailable key information was transparently documented. Additionally, automated checks identified formatting inconsistencies, metadata mapping errors, and numerical anomalies, ensuring a structured and standardized dataset. For verification, we collaborated with historians, archaeologists, and museum curators to review each artifact’s description, cultural attribution, and historical significance. Expert validation ensured that generated textual descriptions were accurate, contextually relevant, and aligned with historical records. This rigorous process enhances the dataset’s credibility, making it a valuable resource for historical research, machine learning, and cultural heritage preservation while ensuring reliable insights into human history. Additional details are presented in Appendix (Sec. D).

3 TimeTravel Benchmark Evaluation

Evaluation Metric: To assess the quality, accuracy, and relevance of our generated textual descriptions, we employed a combination of traditional and advanced metrics. BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) evaluate linguistic fluency and structural similarity, ensuring syntactic alignment with reference texts. METEOR (Banerjee and Lavie, 2005) enhances this by incorporating synonym matching and paraphrasing, improving adaptability to human variations. SPICE (Anderson et al., 2016) assesses semantic accuracy through scene graph analysis, preserving object relationships and cultural context. Additionally, BERTScore (Zhang et al., 2019) offers a deep

	Model	BLEU	METEOR	ROUGE-L	SPICE	BERTScore	LLM-Judge
Closed	GPT-4o-0806 (OpenAI, 2024)	0.1758	0.2439	0.1230	0.1035	0.8349	0.3013
	Gemini-2.0-Flash (Reid et al., 2024)	0.1072	0.2456	0.0884	0.0919	0.8127	0.2630
	Gemini-1.5-Pro (Reid et al., 2024)	0.1067	0.2406	0.0848	0.0901	0.8172	0.2276
	GPT-4o-mini-0718 (OpenAI, 2024)	0.1369	0.2658	0.1027	0.1001	0.8283	0.2492
Open	Llama-3.2-Vision-Inst (Meta AI, 2024)	0.1161	0.2072	0.1027	0.0648	0.8111	0.1255
	Qwen-2.5-VL (Team, 2025)	0.1155	0.2648	0.0887	0.1002	0.8198	0.1792
	Llava-Next (Liu et al., 2024)	0.1118	0.2340	0.0961	0.0799	0.8246	0.1161

Table 2: Performance comparison of various closed and open-source models on our proposed TimeTravel benchmark.

	Model	India	Roman Empire	China	British Isles	Iran	Iraq	Japan	Central America	Greece	Egypt
Closed	GPT-4o-0806	0.2491	0.4463	0.2491	0.1899	0.3522	0.3545	0.2228	0.3144	0.2757	0.3649
	Gemini-2.0-Flash	0.1859	0.3358	0.2059	0.1556	0.3376	0.3071	0.2000	0.2677	0.2582	0.3602
	Gemini-1.5-Pro	0.1118	0.2632	0.2139	0.1545	0.332	0.2587	0.1871	0.2708	0.2088	0.2908
	GPT-4o-mini-0718	0.2311	0.3612	0.2207	0.1866	0.2991	0.2632	0.2087	0.3195	0.2101	0.2501
Open	Llama-3.2-Vision-Inst	0.0744	0.1450	0.1227	0.0777	0.2000	0.1155	0.1075	0.1553	0.1351	0.1201
	Qwen-2.5-VL	0.0888	0.1578	0.1192	0.1713	0.2515	0.1576	0.1771	0.1442	0.1442	0.2660
	Llava-Next	0.0788	0.0961	0.1455	0.1091	0.1464	0.1194	0.1353	0.1917	0.1111	0.0709

Table 3: Analysis of LLM-Judge evaluation of various models in describing archaeological artifacts across civilizations from different geographical locations. Additional comparisons are presented in Appendix (Table 4).

learning-based evaluation of semantic similarity, capturing contextual meaning beyond simple word overlap. LLM-Judge further enhances assessment by evaluating coherence, factual accuracy, and contextual appropriateness.

Results and Analysis: Our evaluation of closed-source and open-source models on the TimeTravel dataset reveals clear differences in their ability to generate historically accurate descriptions (see Table 2). Among closed-source models, GPT-4o-0806 achieved the highest BLEU (0.1758), ROUGE-L (0.1230), SPICE (0.1035), BERTScore (0.8349), and LLM-Judge score (0.3013), indicating superior semantic alignment and contextual richness. However, its lower METEOR score (0.2439) suggests that while it generates highly structured descriptions, they may lack word-level diversity and fluency. GPT-4o-mini-0718, despite scoring slightly lower in BLEU (0.1369) and ROUGE-L (0.1027), outperformed all models in METEOR (0.2658), highlighting its strength in producing more lexically diverse and well-formed outputs. Gemini-2.0-Flash and Gemini-1.5-Pro, while achieving moderate performance across all metrics, demonstrated weaker lexical alignment (BLEU: 0.1072, 0.1067) and semantic coherence (BERTScore: 0.8127, 0.8172), suggesting that they may struggle with historical specificity and structured descriptions. Among open-source models, Qwen-2.5-VL performed the best, achieving higher BLEU (0.1155), METEOR (0.2648), and SPICE (0.1002) compared to its counterparts. These scores indicate a better balance between fluency

and contextual accuracy, making it a strong contender despite being an open-source model. Llama-3.2-Vision-Inst and Llava-Next, however, showed lower SPICE (0.0648, 0.0799) and LLM-Judge scores (0.1255, 0.1161), suggesting difficulties in capturing object details and historical context.

Table 3 presents the LLM-Judge evaluation of models in describing archaeological artifacts across civilizations from different geographic regions. GPT-4o-0806 outperformed other models in describing archaeological artifacts, excelling in regions like the Roman Empire, Iran, Iraq, and Egypt, indicating strong contextual understanding. GPT-4o-mini-0718 and Gemini-2.0-Flash showed strengths in India, Central America, and China, but with some limitations. Among open-source models, Qwen-2.5-VL performed best in Iran, the British Isles, and Egypt, though overall, closed-source models provided more accurate historical descriptions. Additional analysis based on the METEOR score is presented in Appendix (Table 4).

Overall, closed-source models outperform open-source models in generating context-aware descriptions, but ongoing improvements in open-source models highlight opportunities for fine-tuning and dataset expansion. These findings will guide further model enhancements, advancing AI-driven historical analysis and cultural heritage preservation.

4 Conclusion

We present the TimeTravel dataset, a curated collection of historical artifacts from 10 cultural regions, extensively curated by domain experts. We

developed a rigorous data collection, filtering, and verification process, ensuring accuracy and completeness. Using GPT-4o, we generated detailed textual descriptions, making the dataset more accessible and valuable for AI-driven historical research. Our evaluation, using BLEU, METEOR, ROUGE-L, CIDEr, SPICE, BERTScore, and LLM-Judge, showed that closed-source models outperformed open-source alternatives, though open models are rapidly improving. Our analysis highlights the potential of LMMs in bridging gaps in historical records while maintaining academic integrity. By leveraging AI-driven methodologies, this work sets the foundation for advancing cultural heritage preservation and enhancing digital humanities research, ensuring greater accessibility and accuracy in historical documentation.

5 Limitations and Societal Impact

While this research demonstrates the potential of LMMs in enhancing historical documentation, the quality of generated descriptions depends on the completeness and accuracy of the input data. In cases where historical records are fragmented or ambiguous, AI-generated text may lack full contextual depth. Additionally, biases present in training data can influence how models interpret and describe cultural artifacts, necessitating continuous evaluation and expert validation to ensure historical accuracy and cultural sensitivity. Despite these challenges, this research contributes to cultural heritage preservation, educational accessibility, and AI-driven humanities research. By digitizing and enriching historical records, it enables wider public engagement with history, supports museum digitization efforts, and provides a foundation for future advancements in AI-assisted historical analysis, bridging the gap between technology and human expertise in understanding our collective past.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. "Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Badr AlKhamissi, Muhammad N. ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. *Investigating cultural alignment of large language models*. *ArXiv*, abs/2402.13231.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks. *arXiv preprint arXiv:2403.01031*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Yannis Assael, Thea Sommerschild, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, J. Prag, and Nando de Freitas. 2022. *Restoring and attributing ancient texts using deep neural networks*. *Nature*, 603:280 – 283.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Fan Bu, Zheng Wang, Siyi Wang, and Ziyao Liu. 2025. An investigation into value misalignment in llm-generated texts for cultural heritage. *arXiv preprint arXiv:2501.02039*.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *arXiv preprint arXiv:2410.02677*.

Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Husain Salem Abdulla Alharthi, Ines Riahi, Abduljalil Saif, Jorma Laaksonen, Fahad Shahbaz Khan, Salman H Khan, and Rao Muhammad Anwer. 2025. Camel-bench: A comprehensive arabic lmm benchmark. *NAACL*.

Ahmed Heakl, Sara Ghaboura, Omkar Thawkar, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Salman Khan. 2025. *Ain: The arabic inclusive large multimodal model*.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

- Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*.
- Meta AI. 2024. [Llama 3.2: Revolutionizing edge ai and vision with open, customizable models](#).
- Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.
- OpenAI. 2024. [Gpt-4o mini: advancing cost-efficient intelligence](#).
- OpenAI. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Aida Ramezani and Yang Xu. 2023. [Knowledge of cultural moral norms in large language models](#). *ArXiv*, abs/2306.01857.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and et al. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *Preprint*, arXiv:2403.05530.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.
- Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.
- Qwen Team. 2025. [Qwen2.5-vl](#).
- Yingtao Tian, Chikahiko Suzuki, Tarin Clanuwat, Mikel Bober-Irizar, Alex Lamb, and Asanobu Kitamoto. 2020. Kaokore: A pre-modern japanese art facial expression dataset. *arXiv preprint arXiv:2002.08595*.
- Caroline Tully. 2020. British museum. In *Encyclopedia of Global Archaeology*, pages 1618–1620. Springer.
- Michael EW Varnum, Nicolas Baumard, Mohammad Atari, and Kurt Gray. 2024. Large language models based on historical text could offer informative tools for behavioral science. *Proceedings of the National Academy of Sciences*, 121(42):e2407639121.
- Mei Wang and Weihong Deng. 2022. Oracle-mnist: a realistic image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:2205.09442*.
- Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Jinpeng Wan, Haisu Guan, Zhebin Kuang, Lianwen Jin, Xiang Bai, et al. 2024. An open dataset for oracle bone script recognition and decipherment. *arXiv preprint arXiv:2401.15365*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

In this appendix, we provide additional details to support our research, including related work, data statistics, and a comprehensive overview of archaeological samples from various cultures, civilizations, and dynasties. The related work section provides a review of existing research in AI-driven historical text generation, contextualizing our contributions within the broader field. The data statistics section offers a structured breakdown of collected samples, highlighting their geographical distribution and cultural significance. Additionally, the inclusion of archaeological records from diverse historical periods reinforces the depth and diversity of the dataset.

B Related Work

Recent years have seen significant progress in studying cultural representation in AI, particularly in behavioral patterns, food, landmarks, and historical knowledge. However, most works focus on misalignment and biases in AI models or modern cultural trends, rather than positioning artifacts within their historical context and era across ancient civilizations. Meanwhile, studies on cultural inclusion in LLMs highlight the challenges of capturing the contextual and multifaceted nature of culture, emphasizing the limitations of text-based models in representing underrepresented cultures and the need for more robust evaluation methods (Adilazuarda et al., 2024).

Research on cultural influences in AI has increasingly focused on biases and misalignment in language models, particularly how they reflect and perpetuate dominant cultural norms. Early research on cultural biases in LLMs revealed their alignment with Western norms, particularly in moral reasoning, historical narratives, and societal values. Ramezani et al. (2023) analyze how monolingual English language models tend to reflect Western moral norms more strongly than diverse cultural perspectives, limiting their applicability in cross-cultural ethical contexts (Ramezani and Xu, 2023). Tao et al. (2024) further highlight the overrepresentation of Anglo-American and Protestant European values in AI-generated content, often underrepresenting non-Western traditions and belief systems (Tao et al., 2024). Similarly, Bu et al. (2025) explore value misalignment in cultural heritage-related text generation, warning of historical inaccuracies, cultural identity erosion,

and oversimplification of complex narratives, with 65% of the generated content showing significant misalignment (Bu et al., 2025).

To mitigate these biases, several approaches have been proposed. AlKhamissi et al. (2024) introduce Anthropological Prompting, a method that encourages LLMs to reason like cultural anthropologists by incorporating both emic (insider) and etic (outsider) perspectives (AlKhamissi et al., 2024). Similarly, Li et al. (2024) propose CultureLLM, a fine-tuning approach designed to integrate cultural knowledge into LLMs, particularly for low-resource cultures (Li et al., 2024). While these techniques improve cultural alignment, their focus remains on modern cultural settings, leaving gaps in historical artifact contextualization across different time periods.

With the rise of Vision-Language Models (VLMs), cultural research has expanded to multimodal AI, revealing similar biases. Liu et al. (2025) introduce CultureVLM, a model designed to improve cultural understanding in VLMs, highlighting their inability to recognize non-Western cultural symbols, historical artifacts, and traditional gestures (Liu et al., 2025). Their work also presents CultureVerse, a large-scale multimodal dataset covering several cultural concepts, designed to evaluate VLMs' cultural reasoning. However, CultureVerse has a primary focus on modern cultural symbols, traditions, and everyday life. Additionally, Romero et al. (2024) develop CVQA, a multilingual and culturally diverse Visual Question Answering (VQA) benchmark, which reveals that state-of-the-art VLMs struggle with culturally grounded reasoning, particularly in non-Western contexts (Romero et al., 2024). However, these datasets primarily focus on present-day cultural contexts, even when historical artifacts are included, as they are often framed through the lens of modern nations rather than their original civilizations and historical epochs (Liu et al., 2025). This leaves a significant gap in representing artifacts within their authentic temporal and cultural contexts.

Efforts to bridge AI research with historical studies have led to the development of Historical Large Language Models (HLLMs), trained on historical texts to simulate past societies' psychology and value systems (Varnum et al., 2024). These models aim to provide insight into long-term cultural evolution, but their reliance on text-only representations limits their application in multimodal historical

Model	India	Roman Empire	China	British Isles	Iran	Iraq	Japan	Central America	Greece	Egypt
GPT-4o-0806 (OpenAI, 2024)	0.2566	0.2713	0.2324	0.2175	0.2486	0.2428	0.2269	0.2384	0.2441	0.2567
Gemini-2.0-Flash (Reid et al., 2024)	0.2478	0.2603	0.2183	0.2189	0.2432	0.242	0.2256	0.2264	0.2488	0.2588
Gemini-1.5-Pro (Reid et al., 2024)	0.2586	0.2596	0.2198	0.2203	0.2535	0.2524	0.2253	0.2218	0.2551	0.268
GPT-4o-mini-0718 (OpenAI, 2024)	0.2762	0.2731	0.2570	0.2531	0.2660	0.2640	0.2611	0.2741	0.2649	0.2741
Llama-3.2-Vision-Inst (Meta AI, 2024)	0.2128	0.2253	0.1867	0.1917	0.2115	0.2078	0.1944	0.1979	0.2138	0.2182
Qwen-2.5-VL (Team, 2025)	0.2707	0.2815	0.2526	0.2464	0.2607	0.2631	0.2499	0.2587	0.2713	0.2827
Llava-Next (Liu et al., 2024)	0.2482	0.2527	0.2156	0.2192	0.2389	0.2321	0.2207	0.2196	0.2388	0.2427

Table 4: Analysis of METEOR Evaluation of various models in describing archaeological artifacts across civilizations from different geographical regions.

studies. Similarly, Assael et al. (2022) introduce Ithaca, a deep learning model designed to assist historians in restoring, geographically attributing, and dating ancient Greek inscriptions, significantly improving accuracy over traditional methods (Assael et al., 2022). While these works contribute to historical AI, they primarily focus on text-based reconstruction rather than multimodal representations of historical artifacts across civilizations.

TimeTravel fills this gap by providing a 10k historical artifact open-source dataset spanning 10 ancient world regions (prehistoric and historic), offering the first benchmark to evaluate LMMs on temporal-cultural understanding with expert verification. Unlike prior datasets focused on contemporary cultural knowledge, TimeTravel enables AI models to contextualize artifacts within their historical era, ensuring a more accurate representation of civilizations and their material culture. With domain expert verification, the dataset enhances reliability and authenticity, mitigating potential biases and inaccuracies in AI-generated interpretations. By integrating both textual and multimodal perspectives, TimeTravel advances research in historical-cultural AI, enabling AI systems to better understand and reason about artifacts in their original context.

C TimeTravel Samples Regional Distribution

Fig. 4 illustrates the balanced regional distribution of dataset samples based on archaeological provenance. Greece holds the largest share at 18%, followed by multiple regions, including the Roman Empire, China, British Isles, Egypt, Iraq, and Iran, each at 10%. Japan (9%), India (8%), and Central America (5%) contribute smaller yet significant portions. Overall, the dataset ensures diverse cultural representation without dominance by any single region.

Tables 5 to 14 present further details about

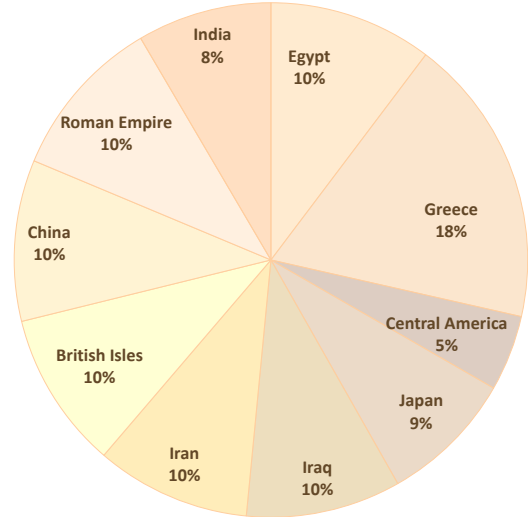


Figure 4: Regional distribution of dataset samples based on their archaeological provenance. Greece holds the largest share at 18%, with a balance-like distribution over regions.

sample counts categorized by region of discovery, section, and cultural affiliation.

The covered areas in our study are ordered as follows:

Tab. 5 → “Roman Empire”, Tab. 6 → “Greece”, Tab. 7 → “British Isles”, Tab. 8 → “Central America”, Tab. 9 → “Egypt”, Tab. 10 → “India”, Tab. 11 → “Iran”, Tab. 12 → “China”, Tab. 13 → “Japan”, and Tab. 14 → “Iraq”,

Place	Roman Empire
Section	Roman
Culture	Samples
Roman Imperial	610
Roman	3
Roman Provincial	436
Total	1049

Table 5: Culture Sample Counts from the Roman Empire.

D TimeTravel Benchmark Examples

Place	Greece		
Section	Greek		
Culture	Sample	Culture	Sample
Greek; Hellenistic; Roman Imperial	4	Hellenistic; Roman Imperial	2
Attic	806	Middle Corinthian	5
Corinthian	41	East Greek; Classical Greek	1
Attic; Classical Greek	47	Transitional Corinthian	1
Middle Corinthian; Late Corinthian; Archaic Greek	7	Classical Greek; Attic	2
Proto-Corinthian	4	Classical Greek; Attic; Archaic Greek	1
Orientalising Period	14	East Greek Archaic II; Archaic Greek	1
Archaic Greek; Classical Greek	1	Attic; Western Greek	1
Archaic Greek	40	East Greek	23
Late Corinthian; Archaic Greek	11	Attic; Archaic Greek	318
Western Greek; Hellenistic	1	Attic; Archaic Greek; Classical Greek	12
Early Corinthian	8	Attic; Classical Greek; Archaic Greek	3
Laconian; Archaic Greek	10	Archaic Greek; East Greek	2
Classical Greek; Corinthian; Hellenistic	1	Rhodian	3
Late Helladic IIIB	2	Greek; Classical Greek	2
Transitional Corinthian; Archaic Greek	1	Early Corinthian; Archaic Greek	3
East Greek; Hellenistic	2	Middle Corinthian; Archaic Greek	11
Late Geometric IIA; Attic	1	East Greek; Orientalising Period	1
Archaic Greek; Attic	8	Late Minoan I; Late Minoan II	1
Late Minoan I	2	Archaic Greek; East Greek; North Ionian	1
Paestan	1	East Greek; Archaic Greek	237
Early Corinthian; Middle Corinthian; Archaic Greek	1	Greek; Hellenistic	2
Archaic Greek; East Dorian	1	Greek	3
Hellenistic	110	Western Greek	5
East Greek; Archaic Greek; Classical Greek	1	Roman; Hellenistic	3
East Dorian; Archaic Greek	2	Classical Greek	38
East Greek; East Dorian; Archaic Greek	11	Boeotian	25
Geometric Greek; Early Proto-Attic	1	Hellenistic; Classical Greek	2
East Greek; South Ionian	1	Geometric Greek	8
Greek; Classical Greek; Hellenistic	5	Hellenistic; Roman	4
Total			1869

Table 6: Culture Sample Counts from Greece (Greek Section).

Place	British Isles
Section	Viking
Culture	Samples
Viking; Carolingian; Late Anglo-Saxon	1
Viking; Early Anglo-Saxon; Mid. Anglo-Saxon	1
Middle Anglo-Saxon Viking; Anglo-Saxon	1
Celtic; Viking	14
Viking; Late Anglo-Saxon	19
Viking; Finno-Ugrian	1
Anglo-Viking	52
Viking	895
Carolingian; Viking	1
Viking; Medieval	1
Late Anglo-Saxon; Viking	1
Viking; Celtic	26
Total	1013

Table 7: Culture Sample Counts from the British Isles (Viking Section).

Place	Central America
Section	Maya
Culture	Samples
Classic Maya; Classic	3
Classic Maya; Late Preclassic Maya	64
Formative (Pre-Classic); Early Classic	8
Maya	
Late Classic Maya	23
Olmec; Maya	1
Classic Maya	275
Preclassic Maya	10
Classic Maya; Late Classic	2
Classic Maya; Olmec	1
Preclassic Maya; Classic Maya	2
Maya	95
Late Classic Maya; Late Classic	4
Total	488

Table 8: Culture Sample Counts from Central America (Maya Section).

Place	Egypt
Section	Ancient Egyptian
Culture	Samples
6 th Dynasty	1
Late Cypriot; 18 th Dynasty	1
26 th Dynasty; Archaic Greek; Punic	1
Late Period; 30 th Dynasty	1
30 th Dynasty; Ptolemaic	15
22 nd Dynasty	69
18 th Dynasty; 19 th Dynasty	2
New Kingdom; 19 th Dynasty; 20 th Dynasty	1
12 th Dynasty	1
26 th Dynasty; Archaic Greek; East Greek; Hellenistic	1
New Kingdom	21
Late Predynastic; 1 st Dynasty	2
25 th Dynasty	7
30 th Dynasty	128
Middle Kingdom	1
Late Period	96
18 th Dynasty; 21 st Dynasty	1
21 st Dynasty	171
19 th Dynasty; 20 th Dynasty	3
20 th Dynasty; 21 st Dynasty	2
26 th Dynasty	257
19 th Dynasty	40
18 th Dynasty	95
1 st Dynasty	7
Ramesside	21
23 rd Dynasty	1
22 nd Dynasty; 23 rd Dynasty	3
26 th Dynasty; 27 th Dynasty	1
20 th Dynasty	25
Late Period; 26 th Dynasty	3
25 th Dynasty; Kushite	1
26 th Dynasty; Punic; Archaic Greek	1
27 th Dynasty	13
25 th Dynasty; 26 th Dynasty	1
Third Intermediate	47
Late Period; Archaic Greek	2
Late Period; Ptolemaic	5
29 th Dynasty	1
New Kingdom; Third Intermediate	1
Ancient Egypt	5
26 th Dynasty; 30 th Dynasty	1
Total	1056

Table 9: Culture Sample Counts from Egypt (Ancient Egyptian Section).

Place	India
Section	Mohenjo-Daro
Culture	Sample Count
Indus Valley Civil.	114
Section	Mauryan
Culture	Sample Count
Mauryan	17
Section	Gupta Dynasty
Culture	Sample Count
Gupta	737
Total	868

Table 10: Culture Sample Counts from India.

Place	Iran
Section	Persian
Culture	Samples
Inju Dynasty	3
Middle Islamic; Seljuq Dynasty; Persian	1
Safavid Dynasty; Mughal Dynasty	1
Persian; Islamic	11
Persian; Late Islamic	3
Samanid Dynasty	27
Safavid Dynasty	395
Timurid Dynasty; Islamic	1
Safavid Dynasty; Post-Medieval	1
Mughal Dynasty; Persian	1
Ilkhanid Dynasty; Persian	3
Turkman Dynasty	3
Early Sasanian; Safavid Dynasty	1
Islamic; Safavid Dynasty	1
Ilkhanid Dynasty	192
Middle Islamic; Persian	6
Islamic; Qajar Dynasty	2
Persian; Safavid Dynasty	1
Safavid Dynasty; Persian; Islamic	2
Mughal Dynasty; Safavid Dynasty	1
Qajar Dynasty	193
Safavid Dynasty; Islamic	4
Persian; Mughal Dynasty	1
Islamic; Persian	2
Timurid Dynasty	35
Persian	108
Total	999

Table 11: Culture Sample Counts from Iran (Persian Section).

Place	China
Section	Tang Dynasty
Culture	Samples
Tang Dynasty; Sui Dynasty	1
Tang Dynasty; Ming Dynasty	3
Tang Dynasty; Ming Dynasty; Jin Dynasty; Yuan Dynasty	1
Tang Dynasty; Song Dynasty	1
Song Dynasty; Tang Dynasty	1
Liao Dynasty; Tang Dynasty	2
Tang Dynasty; Northern Wei Dynasty	1
Six Dynasties; Tang Dynasty	5
Tang Dynasty	1
Northern Qi Dynasty; Sui Dynasty; Tang Dynasty	1
Tang Dynasty; Liao Dynasty	3
Six Dynasties; Sui Dynasty; Tang Dynasty	1
Tang Dynasty; Five Dynasties; Northern Song Dynasty	381
Song Dynasty	
Five Dynasties; Tang Dynasty	4
Tang Dynasty	628
Sui Dynasty; Tang Dynasty	5
Total	1039

Table 12: Culture Sample Counts from China (Tang Dynasty Section).

Place	Japan
Section	Japanese
Culture	Samples
Momoyama Period	6
Genroku Era; Hoei Era	1
Asuka Period	1
Muromachi Period; Momoyama Period	2
Late Kofun; Nara Period	1
Nara Period	12
Middle Kofun	13
Yayoi Period	5
Middle Kofun; Late Kofun	34
Edo Period; Kamakura Period	1
Oei Era	2
Kyowa Era; Oei Era	1
Edo Period; Momoyama Period	1
Jomon Period	16
Kyowa Era	1
Bunka Era	1
Bun' An Era; Bunsei Era	1
Muromachi Period	40
Asuka Period; Nara Period	1
Heian Period	9
Muromachi Period; Momoyama Period; Edo Period	1
Muromachi Period; Buddhist	1
Meiji Era	1
Hakuho Period	1
Showa Era	13
Early Kofun; Middle Kofun	26
Nanbokucho Period	2
Kofun Period; Edo Period	1
Edo Period	24
Kamakura Period; Meiji Era	1
Kofun Period	419
Early Kofun	7
Wado Era	1
Late Kofun	179
Kofun Period; Asuka Period	5
Kamakura Period	26
Nara Period; Edo Period	1
Kofun Period; Nara Period	1
Kamakura Period; Muromachi Period	9
Heian Period; Kamakura Period	1
Total	869

Table 13: Culture Sample Counts from Japan (Japanese Section).

Place	Iraq
Section	Mesopotamian
Culture	Samples
Neo-Assyrian; Late Babylonian	9
Late Babylonian; Assyrian	1
Elamite; Third Dynasty Of Ur	1
Early Dynastic (Middle East)	1
Old Assyrian; Early Bronze Age III	1
Late Uruk	26
Isin-Larsa	3
Neo-Assyrian	406
Uruk	3
Late Uruk; Chalcolithic	1
Middle Babylonian; Neo-Babylonian Dynasty	1
Old Babylonian; Cypriot	1
Late Babylonian	20
Babylonian; Neo-Assyrian	1
Neo-Assyrian; Babylonian	1
Assyrian; Late Babylonian	2
Jemdet Nasr; Proto-Elamite	1
Halaf	38
Assyrian	7
Middle Assyrian	11
Jemdet Nasr	27
Third Dynasty Of Ur; Ubaid	1
Old Babylonian	41
Kassite	4
Babylonian	3
Neo-Babylonian Dynasty	2
Babylonian; Akkadian	1
Old Assyrian	2
Old Babylonian; Third Dynasty Of Ur	1
Ubaid	15
Early Dynastic (Middle East); Akkadian	2
Early Dynastic II	2
Isin-Larsa; Old Babylonian	1
Jemdet Nasr; Akkadian	1
Old Babylonian; Assyrian	1
Akkadian	102
Early Dynastic III; Akkadian	10
Old Babylonian; Old Assyrian	1
Isin-Larsa; Old Babylonian; Kassite	1
Uruk; Jemdet Nasr	3
Early Dynastic II; Early Dynastic III	1
Early Dynastic III	81
Mesopotamian	1
Late Babylonian; Neo-Assyrian	3
Assyrian; Ubaid	1
Third Dynasty Of Ur; Old Babylonian	2
Third Dynasty Of Ur	137
Neo-Assyrian; Phoenician	2
Middle Babylonian	11
Lagash II	1
Third Dynasty Of Ur; Isin-Larsa	4
Total	1000

Table 14: Culture Sample Counts from Iraq (Mesopotamian Section).



```
{
  "id": 4621,
  "Image": "../Documents/2014_11/5_17/e572b5a1_adcb_4691_871c_a3da0123a3fa/preview_01222265_001.jpg",
  "Description": "Silver coin.; Bust of Skandagupta, facing to the right, wearing a decorated cap and collar.; Peacock, with outspread wings and tail, standing facing with head to left. Surrounded by a border of dots.",
  "Production date": "456-467 (about)",
  "Find spot": null,
  "Materials": "silver",
  "Technique": null,
  "Inscription": "Inscription type: inscription Inscription position: reverse Inscription language: Sanskrit Inscription script: Brahmi",
  "Subjects": "king/queen; peacock",
  "Assoc name": "Named in inscription & portrayed: Skandagupta",
  "Culture": "Gupta",
  "Section": "Gupta dynasty",
  "Place": "India"
}
```

for this artifact, we generated a description such as:

```
{
  "description": "This silver coin from the Gupta dynasty, produced approximately between 456-467 CE, features a distinguished portrait of Skandagupta on the obverse. Wearing an ornate cap and collar, Skandagupta is depicted facing right, encapsulating the regal aesthetic of the period. The reverse displays a majestic peacock, wings and tail fully spread, elegantly poised with its head turned to the left, bordered by a circle of dots adding to the coin's decorative appeal. An inscription in Sanskrit, using the Brahmi script, adorns the reverse, providing linguistic and cultural context. This artifact, emblematic of Gupta artistry and power, underscores the dynastic symbolism and reverence for nature characteristic of this era in Indian history."
}
```



```
{
  "id": 7443,
  "Image": "../Documents/2015_9/4_10/b0f2152e_0ab5_4585_a7ca_a50900a611e8/preview_1997_q_1879_a_a.jpg",
  "Description": "Votive object in the form of polished piece of jade with six holes drilled in it.",
  "Production date": null,
  "Find spot": "Excavated/Findspot: Tzimin Kax (?); Excavated/Findspot: Hatzcap Ceel (?); Excavated/Findspot: Cahal Pichik (?)",
  "Materials": "jade",
  "Technique": "polished; drilled",
  "Inscription": null,
  "Subjects": null,
  "Assoc name": null,
  "Culture": "Classic Maya; Late Preclassic Maya",
  "Section": "Maya",
  "Place": "Central America"
}
```

for this artifact, we generated a description such as:

```
{
  "description": "This votive object, crafted by the Classic and Late Preclassic Maya cultures, is a testament to the artistry and spiritual practices of ancient Central America. Made from polished jade, a material revered for its symbolic and aesthetic value, this artifact features six precision-drilled holes, showcasing the sophisticated techniques employed by the Maya artisans. Although specific inscription details are absent, the object's polished surface and drilled holes suggest it may have had ceremonial or religious significance. Unearthed at find spots such as Tzimin Kax, Hatzcap Ceel, and Cahal Pichik, this piece provides valuable insights into the ritualistic and cultural engagements of the Maya civilization."
}
```

Figure 5: This entry represents a silver coin from the Gupta dynasty from India, featuring a distinguished portrait of Skandagupta on the obverse. GPT-4o generated a detailed, context-aware description based on the available metadata, highlighting its craftsmanship, ceremonial significance, and cultural context.

Figure 6: This particular entry represents a polished jade votive object from the Classic and Late Preclassic Maya features six precision-drilled holes, reflecting advanced craftsmanship and likely ceremonial significance. Unearthed at sites like Tzimin Kax, it offers insight into Maya rituals.



Figure 7: Cultural and material diversity of TimeTravel dataset samples across civilizations and historical periods. The dataset includes artifacts from Ancient Egypt, Greece, Mesopotamia, China, and Japan, spanning prehistoric to medieval times. A wide range of materials, including ceramics, metals, and stone, highlights artistic, technological, and societal influences, ensuring a comprehensive representation of historical craftsmanship and cultural heritage.

Roman Imperial / Roman Empire	GPT-4o	Gemini	Qwen	Llava-Next	Ground-Truth
	<p>This image features an ancient Roman coin, possibly a sestertius. The coin has two sides, each with intricate designs typical of Roman craftsmanship. The obverse (front side) shows a profile portrait of a Roman emperor. He is depicted with a laurel wreath on his head, a common symbol of victory and authority. The reverse side shows a seated figure, possibly representing a specific deity or a personification of a virtue. The coin's surface is dark and shows signs of age and wear.</p>	<p>Here's a detailed description of the image. The image shows a detailed view of an ancient Roman coin, likely a sestertius. The coin is circular and features a profile portrait of a Roman emperor on the obverse (front) side. The emperor is wearing a laurel wreath, a symbol of victory and authority. The reverse side of the coin shows a seated figure, possibly representing a deity or a personification of a virtue. The coin is dark and shows signs of age and wear.</p>	<p>The image depicts two sides of an ancient Roman coin, likely a sestertius. The obverse (front) side shows a profile portrait of a Roman emperor, possibly Augustus, wearing a laurel wreath. The reverse side shows a seated figure, possibly a deity or a personification of a virtue. The coin is dark and shows signs of age and wear.</p>	<p>The image presents two ancient Roman coins, each with its own unique design and inscription. The coin on the left is a sestertius, featuring a profile portrait of a Roman emperor. The coin on the right is a denarius, showing a seated figure. Both coins are dark and show signs of age and wear.</p>	<p>The Roman Imperial copper alloy coin, produced between 104-111 CE, displays a finely crafted bust of Emperor Trajan, laureate, facing right, with drapery around his left shoulder. The reverse intricately illustrates the Genius of Aegva Traiana as a ivy-god reclined on rocks from which water gushes, symbolized through an arched grotto supported by two columns, holding a reed in his right hand and resting his left on an urn. Latin inscriptions grace both sides of the coin; the obverse bears the legend 'IMP CAES NERVAE TRAIANO AVG GER DAC P M TR COS V P P' and the reverse is inscribed with 'VS P OR OPTIMO PRINCIPI' along with 'AECVA TRAIANA' and 'VS CV' in the exergues. This coin not only reflects the art and cultural nuances of the Roman Empire but also emphasizes Trajan's engineering prowess and the grandeur of Roman aqueducts. Although its precise findspot remains unknown, it is a testament to the widespread influence of Roman currency across the empire.</p>
	<p>The image shows two sides of an ancient coin, which appears to be heavily corroded and covered with a green patina, typical of aged bronze or copper. The coin is irregular in shape, indicating it might have been minted in ancient times when coinage techniques were less precise than today. The obverse (front) side, a faint impression is visible, though the details are obscured by corrosion. It might depict a figure or symbol, but the exact form is difficult to discern due to wear. The reverse side also shows some faint impressions, possibly of a deity or a figure, but they are also heavily obscured by corrosion and degradation. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>Here's a description of the image, incorporating the what appears to be two sides of an ancient coin or artifact. The object has a rough, irregular shape and is heavily corroded with a green patina. It appears to be a coin or a small figurine, but the details are obscured by the corrosion. The surface is uneven and pitted, with some faint impressions that are difficult to discern. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>The image depicts two ancient coins, likely from the Roman Empire, based on the visible details. The coin on the left shows a profile portrait of a Roman emperor, possibly Augustus, wearing a laurel wreath. The coin on the right shows a seated figure, possibly a deity or a personification of a virtue. Both coins are dark and show signs of age and wear.</p>	<p>The image displays two ancient artifacts, each with a distinct shape and texture. The artifact on the left is a small, irregularly shaped object with a dark greenish-brown patina, indicative of age and exposure to the elements. It has a rough, uneven surface. The artifact on the right is a more rounded, bowl-like object with a similar patina. It has a smooth, curved surface and a small opening at the top. Both artifacts appear to be made of a material that has been heavily weathered and corroded over time.</p>	<p>This copper coin, originating from the Gupta dynasty between 376-415 CE, presents a striking depiction of Chandragupta II. The king is portrayed standing and facing left, adorned with a waistcloth and jewelry, with his right hand raised, signifying authority or blessing. To the right of the king, a diminutive attendant figure holds an exquisitely decorated parasol above him, an accessory emblematic of royalty. The coin's reverse side features the Sanskrit inscription 'SANSKRIT', using the Brahmi script, providing authentic cultural and historical context. The coin's surface shows signs of wear and corrosion, reflecting its age and the environmental conditions it has endured. The intricate details of the king's attire and the attendant's parasol are well-preserved, showcasing the high level of craftsmanship and artistic skill characteristic of the Gupta imperial iconography and their numismatic artistry.</p>
	<p>The image displays an ancient artifact that appears to be a coin or a small figurine. It is heavily corroded and covered with a green patina, typical of aged bronze or copper. The artifact is irregular in shape, suggesting it might have been minted in ancient times when coinage techniques were less precise than today. The obverse (front) side, a faint impression is visible, though the details are obscured by corrosion. It might depict a figure or symbol, but the exact form is difficult to discern due to wear. The reverse side also shows some faint impressions, possibly of a deity or a figure, but they are also heavily obscured by corrosion and degradation. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>Certainly! Here's a detailed description of the image, incorporating the what appears to be two sides of an ancient coin or artifact. The object has a rough, irregular shape and is heavily corroded with a green patina. It appears to be a coin or a small figurine, but the details are obscured by the corrosion. The surface is uneven and pitted, with some faint impressions that are difficult to discern. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>The image shows two sides of an ancient coin, likely from the Roman Empire, based on the visible details. The coin on the left shows a profile portrait of a Roman emperor, possibly Augustus, wearing a laurel wreath. The coin on the right shows a seated figure, possibly a deity or a personification of a virtue. Both coins are dark and show signs of age and wear.</p>	<p>The image shows a small, white ceramic figurine that appears to be a representation of a seated Buddha. The figure is depicted in a meditative pose, with one hand resting on the lap and the other on the knee. The figure is wearing a robe that drapes over the legs and the head. The figurine is made of a light-colored material, possibly terracotta or a similar ceramic. It has a smooth, rounded surface and a small opening at the top. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>This exquisite artifact is a seated figure sculpted from white stoneware, dating back to the Tang dynasty, between 618-907 CE. Modeled with precision, the figure represents Avalokitesvara, also known as Guanyin or Kannon, embodying the compassionate bodhisattva revered across various Buddhist traditions. The choice of white stoneware lends the piece an ethereal presence, reflective of the spiritual purity associated with the figure's portraiture. This artwork exemplifies the sophisticated craftsmanship and religious significance prevalent during the Tang dynasty. It was discovered in China, contributing to our understanding of the period's artistic and cultural achievements.</p>
	<p>The image depicts an intricately crafted gold neck ring, often associated with ancient Nordic cultures. The ring is made of a dark, possibly black, material, likely a metal alloy. It features a complex, repeating pattern of interlocking lines and shapes, creating a dense, textured surface. The ring is shown in a circular view, highlighting its intricate design and craftsmanship. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>Here's a detailed description of the image, incorporating elements of ancient culture. The image shows a detailed view of an ancient artifact, likely a necklace or a similar decorative object. The object is made of a dark material, possibly metal, and features a complex, repeating pattern of interlocking lines and shapes. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>The image depicts an intricately woven bracelet, likely of ancient origin, showcasing a design that is both functional and aesthetically pleasing. The bracelet appears to be made from a single, continuous piece of material, possibly a type of wood or a similar natural material. It features a complex, repeating pattern of interlocking lines and shapes, creating a dense, textured surface. The bracelet is shown in a circular view, highlighting its intricate design and craftsmanship. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>The image shows a gold bracelet with a distinctive pattern of interlocking lines and shapes. The bracelet is made of a dark material, possibly metal, and features a complex, repeating pattern of interlocking lines and shapes, creating a dense, textured surface. The bracelet is shown in a circular view, highlighting its intricate design and craftsmanship. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>This Viking neckring, crafted between the 10th and 11th centuries, exemplifies the artistry and metalwork skills of the era. Composed of twisted silver wires, the neckring combines two elements: expanding terminals that are beaten flat and intricately ornamented with rows of stamped triangles on their upper surfaces, and a functional loop. Such neckrings were likely symbols of wealth and status within Viking society, reflecting both aesthetic preferences and technical prowess. This particular artifact was unearthed as part of the Hålon Moor Hoard, located in the British Isles, providing valuable insights into Viking presence and influence in the region.</p>
	<p>The image features an ancient bronze vessel, possibly a 'Ding' as mentioned in the text. The vessel is dark and shows signs of age and wear. It has a wide, shallow body and a short, flared neck. The surface is covered in a dark patina, and there are some faint impressions or markings on it. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>Here's a detailed description of the image, focusing on the artifact's appearance and potential cultural context. The image shows a detailed view of an ancient artifact, likely a bronze vessel or a similar object. The object is dark and shows signs of age and wear. It has a wide, shallow body and a short, flared neck. The surface is covered in a dark patina, and there are some faint impressions or markings on it. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>The image depicts an intricately woven bracelet, likely of ancient origin, showcasing a design that is both functional and aesthetically pleasing. The bracelet appears to be made from a single, continuous piece of material, possibly a type of wood or a similar natural material. It features a complex, repeating pattern of interlocking lines and shapes, creating a dense, textured surface. The bracelet is shown in a circular view, highlighting its intricate design and craftsmanship. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>The image shows an ancient Chinese bronze vessel, possibly a 'Ding' as mentioned in the text. The vessel is dark and shows signs of age and wear. It has a wide, shallow body and a short, flared neck. The surface is covered in a dark patina, and there are some faint impressions or markings on it. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>This exquisite metal vessel, crafted in the form of a large, wide-mouthed jar or bowl, originates from the Qajar dynasty in Iran. Its surface is intricately decorated with delicate floral motifs and arabesque patterns, showcasing the remarkable craftsmanship and ornamental style characteristic of Persian art during this period. Although the precise production date remains unspecified, the vessel was found in Iran, reflecting the diverse artistic expressions and cultural motifs prevalent during the Qajar era.</p>
	<p>This image features an ancient bronze vessel, possibly a 'Ding' as mentioned in the text. The vessel is dark and shows signs of age and wear. It has a wide, shallow body and a short, flared neck. The surface is covered in a dark patina, and there are some faint impressions or markings on it. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>Here's a detailed description of the image, focusing on the artifact's appearance and potential cultural context. The image shows a detailed view of an ancient artifact, likely a bronze vessel or a similar object. The object is dark and shows signs of age and wear. It has a wide, shallow body and a short, flared neck. The surface is covered in a dark patina, and there are some faint impressions or markings on it. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>The image depicts an intricately woven bracelet, likely of ancient origin, showcasing a design that is both functional and aesthetically pleasing. The bracelet appears to be made from a single, continuous piece of material, possibly a type of wood or a similar natural material. It features a complex, repeating pattern of interlocking lines and shapes, creating a dense, textured surface. The bracelet is shown in a circular view, highlighting its intricate design and craftsmanship. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>The image shows an ancient Chinese bronze vessel, possibly a 'Ding' as mentioned in the text. The vessel is dark and shows signs of age and wear. It has a wide, shallow body and a short, flared neck. The surface is covered in a dark patina, and there are some faint impressions or markings on it. The overall appearance is that of an ancient artifact that has survived through centuries, its original details largely lost to time.</p>	<p>This exquisite metal vessel, crafted in the form of a large, wide-mouthed jar or bowl, originates from the Qajar dynasty in Iran. Its surface is intricately decorated with delicate floral motifs and arabesque patterns, showcasing the remarkable craftsmanship and ornamental style characteristic of Persian art during this period. Although the precise production date remains unspecified, the vessel was found in Iran, reflecting the diverse artistic expressions and cultural motifs prevalent during the Qajar era.</p>

Figure 8: Cross-model comparison of generated descriptions for TimeTravel dataset samples, highlighting variations in detail and accuracy. It illustrates differences in descriptive depth across open- and closed-source models, emphasizing the diversity in interpretative approaches and alignment with the ground truth.