On the Cultural Anachronism and Temporal Reasoning in Vision Language Models

Mukul Ranjan[1] **Prince Jha**[1] **Khushboo Kumari**[2] **Zhiqiang Shen**[1]
[1]MBZUAI, UAE [2]Inception, UAE

https://khushboo0012.github.io/tab-vlm-webpage/

Abstract

Vision-Language Models (VLMs) are increasingly applied to cultural heritage materials, from digital archives to educational platforms. This work identifies a fundamental issue in how these models interpret historical artifacts. We define this phenomenon as *cultural anachronism*, the tendency to misinterpret historical objects using temporally inappropriate concepts, materials, or cultural frameworks. To quantify this phenomenon, we introduce the Temporal Anachronism Benchmark for Vision-Language Models (**TAB-VLM**), a dataset of 600 questions across six categories, designed to evaluate temporal reasoning on 1,600 Indian cultural artifacts spanning prehistoric to modern periods. Systematic evaluations of ten state-of-the-art models reveal significant deficiencies on our benchmark, and even the best model (GPT-5.2) achieves only 58.7% overall accuracy. The performance gap persists across varying architectures and scales, suggesting that cultural anachronism represents a significant limitation in visual AI systems, regardless of model size. These findings highlight the disparity between current VLM capabilities and the requirements for accurately interpreting cultural heritage materials, particularly for non-Western visual cultures underrepresented in training data. Our benchmark provides a foundation for enhancing temporal cognition in multimodal AI systems that interact with historical artifacts. The dataset and code are available in our project page.

1 Introduction

Vision Language Models (VLMs) have demonstrated impressive capabilities in understanding visual content across diverse domains, from natural scene analysis to medical imaging (Radford et al., 2021; Yu et al., 2022; Liu et al., 2023b, 2024; Bai et al., 2025; Li et al., 2025; Liu et al., 2025). Recently, their integration into cultural heritage applications has accelerated rapidly, with deployments spanning digital museum collections, educational platforms, and automated cataloging systems (Li et al., 2024a; Hwang et al., 2025; Trichopoulos, 2023; Arnold and Tilton, 2024; Ghaboura et al., 2025). However, as these models increasingly mediate between the public and cultural artifacts, we identify a critical gap: their tendency to interpret historical artifacts through inappropriate temporal lenses, a phenomenon we define as **`cultural anachronism`**.

Cultural anachronism in VLMs represents the systematic misattribution of concepts, techniques, materials, or interpretive frameworks to artifacts from time periods where they did not exist or were not culturally relevant. For example, when a VLM describes a sculpture from the 3rd century BCE using artistic vocabulary that emerged in 19th century European contexts, attributes manufacturing techniques to ancient pottery that were only developed millennia later, or interprets religious iconography through contemporary rather than period-appropriate symbolic frameworks (Liang et al., 2022; Bhatia et al., 2024; Gallegos et al., 2024; Ko et al., 2023). Such temporal-conceptual displacement poses a significant challenges for applications in museum digitization, educational technologies, and cultural heritage preservation, where accuracy in historical representation is paramount (Siliutina et al., 2024; Hwang et al., 2025; Trichopoulos, 2023; Arnold and Tilton, 2024). Unlike general object recognition errors that may cause minor inconveniences, anachronistic interpretations fundamentally misrepresent the cultural and historical significance of artifacts, potentially reinforcing colonial or contemporary biases in historical narratives and undermining the integrity of cultural heritage documentation (Birhane et al., 2021; Thylstrup, 2022).

The challenge of temporal reasoning in VLMs extends beyond simple date recognition to encompass complex understanding of technological evo-

Findings of the Association for Computational Linguistics: ACL 2026, pages 30294–30310

lution, artistic development, and cultural context across historical periods. While existing benchmarks evaluate VLMs on contemporary visual understanding tasks (Li et al., 2025; Chang et al., 2024), none systematically assess their ability to maintain temporal coherence when interpreting historical materials. This gap is particularly substantial given the increasing reliance on AI systems for cultural heritage applications, where the stakes of misinterpretation extend far beyond technical performance metrics to encompass cultural representation, educational accuracy, and preservation of historical knowledge. The problem is compounded by the fact that VLMs are predominantly trained on contemporary visual data, creating inherent biases toward modern interpretive frameworks when encountering historical artifacts.

To address this gap, we introduce the Temporal Anachronism Benchmark for Vision–Language Models (**TAB-VLM**), a comprehensive dataset comprising 600 carefully curated questions across six distinct evaluation categories, utilizing 1,600 artifacts spanning prehistoric to modern Indian history. Figure 1 provides an overview of the artifact collection and curation pipeline used to construct TAB-VLM. Our benchmark specifically targets the temporal reasoning capabilities required for accurate historical artifact interpretation, including period attribution, chronological sequencing, anachronism detection, manufacturing technique identification, material availability assessment, and cultural context understanding. The selection of Indian cultural heritage as our focus domain provides several advantages: it encompasses a vast temporal range from prehistoric Indus Valley artifacts to contemporary works, represents diverse artistic traditions and technological developments, and offers rich documentation that enables precise temporal categorization while highlighting the global importance of non-Western cultural heritage in AI evaluation frameworks.

Our systematic evaluation approach addresses three fundamental research questions that illuminate the scope and nature of cultural anachronism in current VLMs:

1. **RQ1:** To what extent do current state-of-the-art VLMs exhibit cultural anachronism when interpreting historical artifacts, and does this vary across models of different architectures and scales?

2. **RQ2:** Which aspects of temporal reasoning (e.g., chronological ordering, manufacturing technique identification, material appropriateness) present the greatest challenges for VLMs, and what patterns emerge in their anachronistic interpretations?

3. **RQ3:** How does performance on temporal reasoning tasks correlate with general visual understanding abilities, and what implications does this have for the development of VLMs that avoid cultural anachronism?

We evaluate ten state-of-the-art models including both proprietary models (GPT-5.2 (OpenAI, 2025b), GPT-4o (Hurst et al., 2024), GPT-4o-mini (Hurst et al., 2024)) and leading open-source alternatives (Qwen2-VL (Wang et al., 2024), Qwen2.5-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025) series) on our benchmark, and provide the first quantitative characterization of cultural anachronism prevalence and patterns in VLMs. Our findings reveal limitations in the historical reasoning capabilities of current models despite their demonstrated proficiency in general visual understanding tasks, with implications extending beyond technical performance to encompass broader questions of cultural competence, responsible AI deployment, and the preservation of accurate historical knowledge in digital contexts. This work establishes cultural anachronism as a notable evaluation dimension for multimodal AI systems and provides a foundation for developing more temporally-aware and culturally sensitive VLMs.

2 Related Work

Temporal reasoning remains a persistent challenge for AI systems, with benchmarks like TRAM (Wang and Zhao, 2023), Test-of-Time (Fatemi et al., 2024), TemporalVQA (Imam et al., 2025), SpookyBench (Upadhyay et al., 2025) and TimeBench (Chu et al., 2023) revealing consistent deficiencies in time-based understanding despite advances in model architecture (Jain et al., 2023). Recent approaches including self-critique methods (Su et al., 2024), explainable frameworks (Yuan et al., 2024), video-based temporal reasoning (Ko et al., 2023; Chen et al., 2024), and multimodal temporal-causal evaluation (Wang et al., 2025; Padlewski et al., 2024) show only modest improvements, with models particularly struggling to maintain period-appropriate perspectives (Underwood et al., 2025). Cultural understanding presents parallel challenges, with

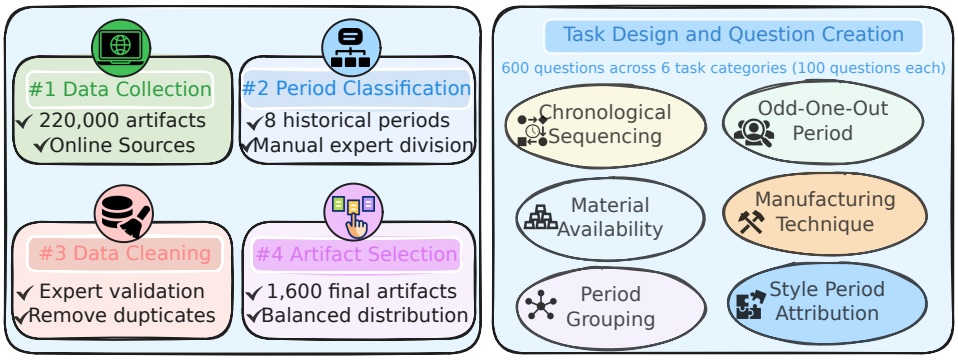

Figure 1: Data collection and processing pipeline showing systematic reduction from 220,000 initial artifacts to 1,600 curated items through expert validation and balanced selection.

research revealing significant gaps in models' cultural knowledge (Li et al., 2024a; Rao et al., 2024; Chiu et al., 2024; Liu et al., 2023a; Hwang et al., 2023), particularly for non-Western traditions and across temporal contexts. Multimodal evaluations (Nayak et al., 2024; Liu et al., 2025; Kannen et al., 2024) demonstrate frequent cultural inaccuracies in VLMs, while frameworks for cross-cultural alignment (Kharchenko et al., 2024; Li et al., 2024b; Fung et al., 2024) remain limited in addressing temporal evolution of cultural expressions.

The intersection of temporal and cultural reasoning becomes important in cultural heritage applications, where AI systems must navigate both historical context and cultural representation. While classification systems for cultural heritage (Hwang et al., 2025) and heritage search interfaces (Trichopoulos, 2023; Arnold and Tilton, 2024) show promise, they struggle with temporal nuance and dynamic cultural evolution (Adilazuarda et al., 2024). Recent benchmarks for historical artifacts (Ghaboura et al., 2025) begin addressing these challenges, but they do not explicitly operationalize cultural–temporal reasoning as an evaluation target, nor do they expose anachronistic failure modes tied to historical artifacts. Temporal reasoning capabilities require deeper examination given the importance of accurate representation in digital cultural preservation (Siliutina et al., 2024). These challenges connect to broader representational biases (Birhane et al., 2021; Thylstrup, 2022; Gallegos et al., 2024) that particularly affect cultural representation, necessitating context-specific approaches (Sambasivan et al., 2021) for artifacts potentially underrepresented in training data. Existing surveys (Li et al., 2025; Chang et al., 2024) document evaluation approaches for VLMs, but do not capture this dimension, motivating the need for targeted benchmarks such as TAB-VLM.

3 TAB-VLM

Our benchmark TAB-VLM focuses on Indian cultural heritage. We curate a diverse dataset of artifacts and construct temporal reasoning tasks that uncover anachronistic responses across six dimensions: (1) Chronological Sequence, (2) Period Intrusion Detection (Odd-One-Out Period), (3) Material Availability, (4) Manufacturing Technique, (5) Period-Based Grouping, and (6) Style Period Attribution. TAB-VLM offers a structured framework for assessing historical awareness in multimodal models.

3.1 Task Design and Evaluation Framework

We construct 600 multiple-choice questions distributed equally across six different temporal reasoning task categories, with 100 questions per category, for the TAB-VLM benchmark. We construct each question using task-specific natural language prompt templates (provided in Appendix B), paired with one or more artifact images. We now describe each of the six task categories in detail.

Chronological Sequencing. The chronological sequencing task evaluates fundamental temporal ordering capabilities by requiring models to arrange artifacts from different periods in correct chronological sequence. Prompt template: "<image1>, <image2>, <image3>, <image4> These are the 4 historical artifacts from different time periods. Please arrange these artifacts in chronological order from oldest to newest." *Task structure:* 4 artifacts with one correct ordering from 24 possible permutations.

Odd-One-Out Period. This task assesses a

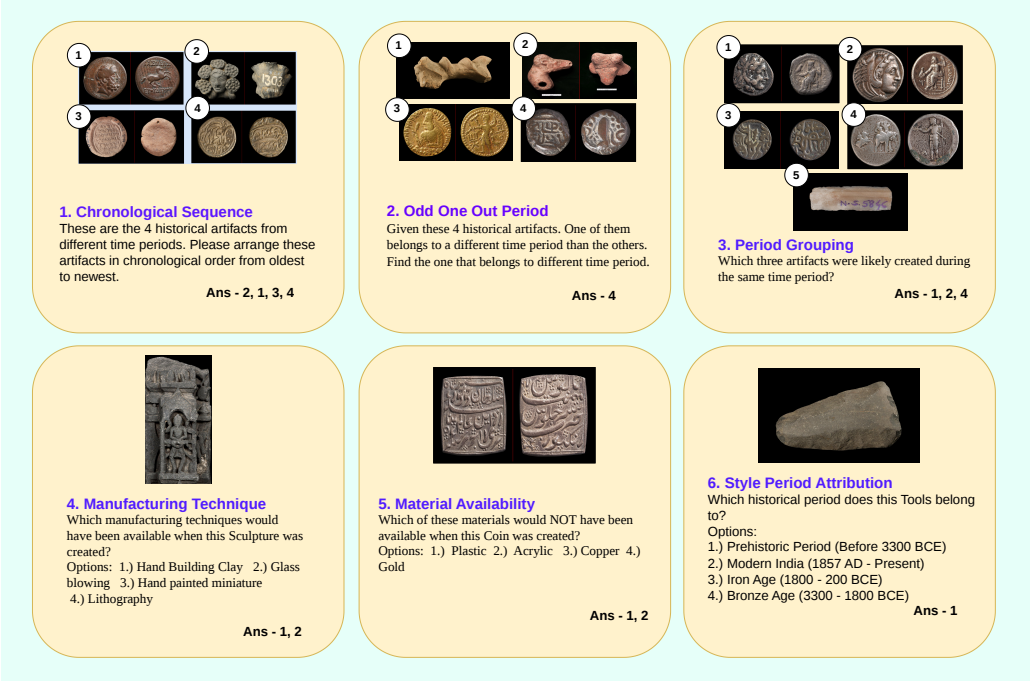

Figure 2: Examples of the six task types in our benchmark: chronological sequencing, odd-one-out period detection, material availability, manufacturing technique, period grouping, and style-period attribution.

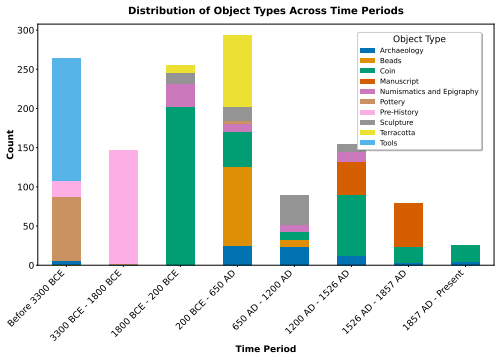

Figure 3: Distribution of object types across time periods indicating prevalence of terracotta artifacts in prehistoric eras and manuscripts during the Early Modern period.

model's sensitivity to temporal anomalies by presenting a group of artifacts, one of which belongs to a different historical period than the others. The model must identify the artifact that does not fit, based on subtle visual cues that indicate differences in era or context. Prompt template: "<image1>, <image2>, <image3>, <image4>, Given these 4 historical artifacts one of them belongs to a different time period than the others. Find the one that belongs to different time period." *Task structure:* 4 options with one correct answer.

Material Availability. This task evaluates a model's understanding of historical material by asking it to identify materials that would not have been available during the artifact's period. It tests knowledge of technological timelines to detect anachronistic material associations. Prompt template: "<image>. Which of the following materials would NOT have been available when this artifact was created?" *Task structure:* 4 material options with 1-2 anachronistic materials (exact set match required).

Manufacturing Technique. This task tests a model's understanding of historical production methods by asking which manufacturing techniques would have been available during the time the artifact was created. It evaluates temporal knowledge of technological capabilities and craftsmanship. Prompt template: "<image>. Which manufacturing techniques would have been available when this artifact was created?" *Task structure:* 4 technique options with 1-2 anachronistic techniques (exact set match required).

Period Grouping. This task examines pattern recognition across stylistic and material elements by requiring identification of contemporaneous artifacts, evaluating models' ability to recognize coherent aesthetic traditions and technological signatures characteristic of specific historical periods. Prompt template: "<image1>, <image2>, <image3>, <image4>, <image5> Which three artifacts were likely created during the same time period?" *Task structure:* 5 artifacts where 3 belong to the same period

Model	Overall Accuracy	Style-Period Attribution	Chronological Sequence	Manufacturing Technique	Material Availability	Period Grouping	Odd-One-Out Period
Proprietary Models							
GPT-5.2	**58.7 ± 1.3**	**65.0 ± 4.1**	**37.2 ± 0.6**	**56.0 ± 1.2**	**92.1 ± 1.3**	**45.2 ± 1.3**	**57.1 ± 1.2**
GPT-5-mini	51.2 ± 0.4	61.0 ± 0.3	32.3 ± 2.1	52.1 ± 1.1	88.0 ± 1.2	36.2 ± 0.8	49.1 ± 1.0
GPT-4o	50.4 ± 0.2	60.3 ± 1.2	30.7 ± 0.0	49.8 ± 1.5	85.3 ± 0.9	32.3 ± 1.3	45.3 ± 1.2
GPT-4o-mini	47.2 ± 0.7	45.0 ± 1.5	34.0 ± 0.7	39.0 ± 1.2	89.0 ± 1.3	28.0 ± 1.5	48.0 ± 0.6
Open-Source Models							
Qwen2-VL-7B	32.8 ± 1.4	52.7 ± 0.9	16.7 ± 0.2	16.7 ± 1.3	70.7 ± 0.6	10.0 ± 1.2	29.3 ± 1.2
Qwen2.5-VL-7B	42.6 ± 1.1	58.8 ± 0.4	14.0 ± 0.3	41.5 ± 0.9	69.0 ± 0.4	22.0 ± 0.6	51.0 ± 0.8
Qwen2.5-VL-3B	29.0 ± 2.1	52.5 ± 1.1	11.5 ± 1.3	40.0 ± 1.7	25.4 ± 0.6	13.3 ± 1.8	30.4 ± 1.1
Qwen2-VL-2B	19.7 ± 0.5	47.5 ± 0.5	1.3 ± 1.8	10.5 ± 0.7	15.2 ± 0.2	8.1 ± 1.3	25.7 ± 0.4
InternVL3-8B	36.2 ± 1.5	53.0 ± 0.4	9.0 ± 0.6	41.0 ± 1.3	59.0 ± 0.3	16.0 ± 0.2	39.0 ± 0.4
InternVL3-2B	30.5 ± 0.1	52.0 ± 0.5	8.0 ± 0.5	42.0 ± 0.6	45.0 ± 1.3	8.0 ± 0.1	28.0 ± 1.2
Baseline							
Random	12.8	25.0	4.2	6.25	6.25	10.0	25.0

Table 1: Performance comparison of VLMs on TAB-VLM benchmark with 100 questions per category. All values represent percentages (mean ± standard deviation across five runs). Best results in bold.

(one correct combination from 10 possible).

Style-Period Attribution. This task provides assessment of temporal classification capabilities by requiring models to match artifacts with their correct historical periods from multiple options. It tests models' understanding of visual characteristics, artistic conventions, and cultural expressions associated with specific temporal periods. Prompt template: "<image> Which historical period does this artifact belongs to?" *Task structure:* 4 period options with one correct answer.

In all the prompt templates above, <image> denotes the visual input of a specific artifact. **Modality clarification:** Models receive only visual input (artifact images) during evaluation; no textual metadata about artifacts is provided. Figure 2 illustrates examples for each task, showing the image, question, and answer triplet. Random baseline performance for each task type is reported in Table 1.

3.2 Dataset Construction and Curation

Our dataset was constructed through a rigorous four-stage pipeline, beginning with the collection of approximately 220,000 artifacts from online repositories (Sahapedia, 2025). As shown in Table 2 these artifacts were then classified into eight distinct historical periods, based on the taxonomy outlined in (McLeod, 2015): Prehistoric Period (before 3300 BCE), Bronze Age (3300–1800 BCE), Iron Age (1800–200 BCE), Classical Period (200–650 AD), Early Medieval Period (650–1200 AD), Late Medieval Period (1200–1526 AD), Early Modern Period (1526–1857 AD), and Modern India (1857 AD–present). To construct 600 multiple-choice questions for the TAB-VLM benchmark, we randomly sampled images from the corresponding historical periods.

Historical Period	Time Range	Artifacts
Prehistoric Period	Before 3300 BCE	276
Bronze Age (Indus Valley Civilization)	3300 – 1800 BCE	150
Iron Age	1800 – 200 BCE	259
Classical Period	200 – 650 AD	310
Early Medieval Period	650 – 1200 AD	139
Late Medieval Period	1200 – 1526 AD	155
Early Modern Period	1526 – 1857 AD	137
Modern India	1857 AD – Present	174
Total	**–**	**1,600**

Table 2: Distribution of Artifacts Across Historical Periods in the TAB-VLM Benchmark

We implement a data cleaning process that included expert validation by the authors of this paper. This process involved eliminating duplicates, removing artifacts with ambiguous or disputed dating, and filtering out items lacking sufficient visual detail for reliable temporal assessment. As a result, the corpus was significantly reduced while maintaining strict quality standards for historical accuracy. The final cleaned dataset consists of 1,600 artifacts.

The final dataset exhibits a temporal distribution reflecting the archaeological record of Indian cultural heritage, with prehistoric periods containing 276 artifacts, representing the rich material culture of early Indian civilizations (Table 2). Medieval periods are well-represented with 139-155 artifacts each, while more recent periods contain

fewer items due to different preservation patterns and the scope of museum collections. The distribution of object types reveals clear temporal patterns: terracotta artifacts dominate prehistoric collections, reflecting early ceramic traditions; manuscripts become prominent during the Late Medieval Period, showcasing court literature and illuminated texts; and stone sculptures span multiple eras, particularly flourishing during the Classical Period (Figure 3). Material composition demonstrates technological evolution across Indian history, with stone artifacts prevalent in prehistoric periods, copper and bronze emerging in ancient times, and paper becoming prominent in medieval and modern India (Figure 4). This material diversity provides crucial ground truth for evaluating models' understanding of technological development timelines and material availability across different historical periods.

4 Experimental Results

4.1 Experimental Setup

We conduct our experiments using ten different VLMs, comprising six open-source and four proprietary models to provide comprehensive coverage of the current VLM landscape. The open-source models include instruction-tuned variants from the Qwen2-VL (Wang et al., 2024) series with 2B and 7B parameters, the Qwen2.5-VL (Bai et al., 2025) series with 3B and 7B parameters, and InternVL3 (Zhu et al., 2025) models with 2B and 8B parameters, representing state-of-the-art open-source capabilities across different model scales. As proprietary baselines, we evaluate OpenAI's GPT-5.2 (OpenAI, 2025b), GPT-5-mini (OpenAI, 2025a), GPT-4o (Hurst et al., 2024) and GPT-4o-mini (Hurst et al., 2024), which represent the current frontier in commercial VLM performance. All open-source model experiments are conducted on a single NVIDIA A100 GPU using the default hyperparameters from the HuggingFace (Wolf et al., 2019) implementation with default hyperparameters. All models are evaluated using the same prompt templates (prompt provided in Appendix B) and evaluation pipeline to ensure fair comparison, with each question processed independently and model responses parsed using consistent answer extraction rules tailored to each question type.

4.2 Evaluation Metric

We evaluate model performance using accuracy as the primary metric, calculated as the proportion of correctly answered questions within each task category and overall. For single-choice questions (Style-Period Attribution, Odd-One-Out Period Detection), accuracy represents exact match between predicted and ground truth answers. For multi-choice questions (Manufacturing Technique, Material Availability), we require exact set matching where all correct options must be selected and no incorrect options chosen. For sequence-based tasks (Chronological Sequencing), accuracy measures exact ordering match, while for grouping tasks (Period Grouping), we require precise identification of all three artifacts from the same historical period. Standard deviation is computed across multiple evaluation runs to assess result stability.

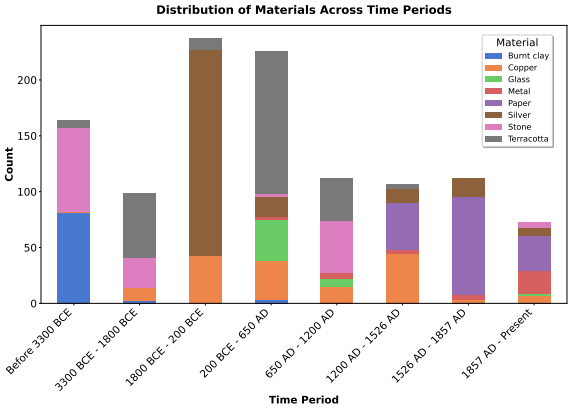

Figure 4: Distribution of materials across historical time periods showing predominance of stone artifacts in prehistoric periods and silver in medieval India.

4.3 Results and Analysis

Our experiment reveals several significant problems in temporal reasoning capabilities of VLMs when interpreting historical artifacts. Table 1 presents accuracy results for each model across all task categories, demonstrating substantial performance gaps between proprietary and open-source models. GPT-5.2 achieved the highest overall accuracy at 58.7%, establishing clear superiority across most task categories and representing the current state-of-the-art for temporal reasoning in cultural heritage contexts. The model demonstrated exceptional performance in Material Availability detection (92.1%) and Style-Period Attribution (65.0%), indicating robust understanding of technological constraints and stylistic recognition capabilities that enable identification of anachronistic materials and period-appropriate artistic elements. However, even this best-performing model exhibited limitations in temporal ordering tasks,

achieving only 37.2% accuracy in Chronological Sequencing and 45.2% in Period Grouping, revealing gaps in understanding temporal relationships between artifacts across different historical periods. GPT-5-mini (OpenAI, 2025a) and GPT-4o-mini achieved competitive overall performance at 51.2% and 47.2% respectively. GPT-4o-mini notably surpasses GPT-4o in Material Availability (89.0% vs 85.3%) and Odd-One-Out Period Detection (48.0% vs 45.3%), suggesting that cost-effective models can achieve comparable or superior performance on specific temporal reasoning subtasks. Among open-source alternatives, Qwen2.5-VL-7B emerged as the strongest performer with 42.6% overall accuracy, demonstrating particularly strong capabilities in Style-Period Attribution (58.8%) and achieving the highest open-source performance in Odd-One-Out Period Detection (51.0%), indicating effective training on temporal pattern recognition despite being substantially smaller than proprietary models. InternVL3-8B achieved 36.2% overall accuracy with balanced performance across most categories, while Qwen2-VL-7B reached 32.8% overall accuracy but showed pronounced weaknesses in Manufacturing Technique recognition (16.7%) and Period Grouping (10.0%), suggesting specific limitations in understanding technological evolution and temporal clustering. Smaller parameter models consistently underperformed, with Qwen2.5-VL-3B achieving 29.0% accuracy and Qwen2-VL-2B reaching only 19.7%, though both maintained reasonable Style-Period Attribution capabilities (52.5% and 47.5% respectively), indicating that certain temporal reasoning capabilities may be robust to model scale reduction. Across all evaluated models, Material Availability emerged as the most accessible task category, with six of ten models achieving above 45% accuracy and both GPT variants exceeding 85%, suggesting that models can effectively identify anachronistic material usage through learned associations between materials and historical periods. Conversely, Chronological Sequencing proved universally challenging, with no model exceeding 37.2% accuracy and most open-source models performing below 20%, indicating problems in understanding temporal progression and artifact dating relationships. Period Grouping similarly challenged all models, with performance ranging from 8.0% to 45.2%, revealing difficulties in recognizing shared temporal characteristics across different artifact types and cultural contexts. Notably, the relationship between model scale and temporal reasoning performance proved complex and non-linear, as evidenced by Qwen2.5-VL-3B achieving comparable or superior performance to larger models in several categories, while InternVL3-8B showed only marginal improvements over InternVL3-2B (36.2% vs 30.5%), suggesting that architectural innovations, training methodology, and dataset composition may be more influential factors than raw parameter count for developing effective temporal reasoning capabilities in VLMs.

5 Discussion

Our evaluation reveals fundamental limitations in current VLMs' ability to reason temporally about historical artifacts. We address our three research questions through quantitative results and qualitative error analysis, providing diagnostic insights into the nature of cultural anachronism in multimodal AI systems.

5.1 Extent of Cultural Anachronism Across Model Architectures

Addressing RQ1, current state-of-the-art VLMs exhibit substantial cultural anachronism when interpreting historical artifacts, with even the best-performing model (GPT-5.2) achieving only 58.7% overall accuracy, a modest improvement over the random baseline of 12.8%. This 45.9 percentage point gap, while substantial, reveals that even frontier models struggle significantly with temporal reasoning. The consistency of poor performance across different model architectures and scales suggests that cultural anachronism represents a fundamental limitation rather than a scaling problem. Notably, GPT-4o's marginal advantage over GPT-4o-mini (50.4% vs 47.2%), and Qwen2.5-VL-3B's competitive performance with much larger models in several categories, indicate that architectural innovations and training methodology may be more critical than raw parameter count. These findings contrast sharply with VLMs' generally strong performance on contemporary visual understanding benchmarks, suggesting that temporal reasoning about historical contexts requires fundamentally different capabilities than those emphasized in current training paradigms.

5.2 Task-Specific Challenges and Error Patterns

Addressing RQ2, we observe pronounced variations in performance across task categories that reveal distinct failure modes.

Chronological Sequencing emerged as the most challenging task, with the best model achieving only 37.2% accuracy compared to a random baseline of 4.2%. Qualitative analysis reveals systematic errors: models frequently confuse adjacent historical periods (e.g., Classical vs. Early Medieval) and show a recency bias, often placing more weathered artifacts as "older" regardless of their actual period. For instance, GPT-5.2 incorrectly ordered a well-preserved Bronze Age Indus seal after a weathered Classical period sculpture, suggesting reliance on surface condition rather than stylistic or technological markers. **Material Availability** proved most accessible (GPT-5.2: 92.1% vs. random baseline: 6.25%), but analysis of the 7.9% error rate reveals concerning patterns. Models occasionally attribute modern materials (synthetic dyes, industrial alloys) to medieval artifacts, or fail to recognize that certain precious metals became available through trade routes in specific periods. This suggests models may rely on general historical knowledge about material development rather than nuanced understanding of regional technological timelines. **Style-Period Attribution** showed intermediate difficulty (GPT-5.2: 65.0% vs. random baseline: 25.0%). Error analysis reveals systematic confusion between adjacent periods, with 68% of misclassifications occurring within ±1 period boundary. Models particularly struggle with transitional artifacts showing mixed stylistic elements, and exhibit a tendency to over-classify artifacts into later periods when visual details are ambiguous, a potential consequence of training data skew toward more recent, better-documented artifacts. **Period Grouping** challenged all models (best: 45.2%; random: 10.0%), with failures revealing an inability to recognize coherent technological and stylistic signatures across different artifact types. Models often group visually similar artifacts (e.g., all terracotta items) regardless of period, suggesting that superficial feature matching overrides temporal reasoning. These patterns indicate that VLMs perform better on tasks solvable through memorized factual associations (material-to-period mappings) than those requiring genuine temporal cognition and visual pattern recognition across stylistic evolution.

5.3 Implications for Development and Deployment

Addressing RQ3, the disconnect between VLMs' strong general visual capabilities and poor temporal reasoning reveals fundamental gaps in current training paradigms. Visual feature extraction enabling modern object detection does not transfer effectively to historical contexts requiring period-appropriate interpretive frameworks.

Risks for Cultural Heritage Applications and Directions for Mitigation. VLMs deployed in museum digitization or educational platforms may systematically misrepresent artifacts, potentially perpetuating colonial-era anachronistic interpretations or erasing nuanced historical developments. The observed issues are particularly problematic for non-Western cultural heritage. Our findings also suggest several avenues for improvement: (1) *Training data curation*: Explicitly include temporally annotated historical corpora with diverse cultural contexts and period coverage; (2) *Specialized objectives*: Incorporate contrastive learning tasks that require distinguishing adjacent periods and recognizing anachronistic material-technique combinations; (3) *Evaluation protocols*: Systematically assess cultural anachronism before deployment, particularly for applications involving non-Western heritage.

6 Conclusion

We introduce TAB-VLM, the first benchmark to systematically evaluate cultural anachronism in VLMs through 600 questions across 1,600 Indian cultural artifacts. Our evaluation of ten state-of-the-art VLMs reveals widespread temporal reasoning deficiencies, with even the best model (GPT-5.2) achieving only 58.7% overall accuracy and overall struggle in chronological sequencing tasks ($\leq 38\%$ accuracy). The consistency of poor performance across different architectures and scales indicates that cultural anachronism represents a fundamental problem in current VLM training paradigms rather than a scaling problem. These findings have several important implications for cultural heritage applications, where anachronistic interpretations risk misrepresenting historical artifacts and perpetuating biased cultural narratives. Our work establishes cultural anachronism as an essential evaluation dimension for multimodal AI and provides a framework for developing temporally-aware systems capable of respectful engagement with cultural heritage materials. We hope that TAB-VLM will catalyze further research into incorporating temporal reasoning and cultural sensitivity into VLM training objectives, particularly for underrepresented non-Western heritage.

7 Limitations and Future Work

Our evaluation is constrained by several methodological limitations that affect generalizability. The benchmark focuses exclusively on Indian cultural artifacts, limiting cross-cultural applicability and potentially missing anachronistic patterns specific to other cultural contexts. Our visual-only evaluation approach excludes textual metadata that would typically accompany artifacts in real deployments. Expert annotations, though rigorously validated, reflect particular scholarly perspectives and discrete temporal classifications that may not accommodate transitional or culturally mixed artifacts. Additionally, our model selection represents only a subset of available VLMs, and the rapid pace of development means newer architectures may exhibit different anachronistic patterns than those evaluated here.

The modest performance differences across model scales indicate that simply scaling existing approaches is insufficient. Apart from improving the dataset discussed above, future work should also investigate model development techniques such as whether fine-tuning on temporally annotated historical datasets can reduce anachronism, and whether incorporating explicit temporal reasoning signals (e.g., stratified period embeddings) during training improves performance. Additionally, cross-cultural validation on artifacts from multiple cultural traditions would clarify whether the observed patterns generalize or reflect culture-specific training data gaps.

8 Use of Language Models

Large language models were used in a limited capacity to assist with minor editing and polishing of the manuscript. The use of LLMs was strictly limited to text formatting and grammatical corrections; no AI tools were used to generate scientific ideas, interpret results, or formulate the core arguments of this work. All technical content, experimental design, results, and conclusions were produced, verified, and finalized by the authors.

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*.

Taylor Arnold and Lauren Tilton. 2024. Explainable search and discovery of visual cultural heritage collections with multimodal large language models. *arXiv preprint arXiv:2411.04663*.

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

Gagan Bhatia, MingZe Tang, Cristina Mahanta, and Madiha Kazi. 2024. Datelogicqa: Benchmarking temporal biases in large language models. *arXiv preprint arXiv:2412.13377*.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Frank Wang. 2024. Rextime: A benchmark suite for reasoning-across-time in videos. In *Advances in Neural Information Processing Systems*, volume 37, pages 28662–28673.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and 1 others. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *arXiv preprint arXiv:2410.02677*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2023. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *arXiv preprint arXiv:2311.17667*.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. *arXiv preprint arXiv:2406.09170*.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Sara Ghaboura, Ketan More, Ritesh Thawkar, Wafa Alghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Time travel: A comprehensive benchmark to evaluate lmms on historical and cultural artifacts. *arXiv preprint arXiv:2502.14865*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*.

Hyerin Hwang, Chan-Woo Park, Hee-Kwon Kim, and Jae-Ho Lee. 2025. Cats: cultural-heritage classification using llms and distribute model. *npj Heritage Science*, 13(1):76.

Mohamed Fazli Imam, Chenyang Lyu, and Alham Fikri Aji. 2025. Can multimodal llms do visual temporal understanding and reasoning? the answer is no! *arXiv preprint arXiv:2501.10674*.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *arXiv preprint arXiv:2407.06863*.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. 2023. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747*.

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

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *arXiv preprint arXiv:2405.15145*.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023a. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

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*.

John McLeod. 2015. *The History of India*. Bloomsbury Publishing USA.

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. 2025a. Gpt-5 system card. Published Aug 13, 2025; system card for the GPT-5 model family.

OpenAI. 2025b. Update to gpt-5 system card: Gpt-5.2. Published Dec 11, 2025; update describing GPT-5.2 model family.

Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, and 1 others. 2024. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. *arXiv preprint arXiv:2405.02287*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Sahapedia. 2025. Museums of india. https://www.museumsofindia.org/. Accessed: 2025-08-02.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 315–328.

Iryna Siliutina, Olena Tytar, Marina Barbash, Nataliia Petrenko, and Larysa Yepyk. 2024. Cultural preservation and digital heritage: challenges and opportunities. *Amazonia Investiga*, 13(75):262–273.

Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. 2024. Timo: Towards better temporal reasoning for language models. *arXiv preprint arXiv:2406.14192*.

Nanna Bonde Thylstrup. 2022. The ethics and politics of data sets in the age of machine learning: Deleting traces and encountering remains. *Media, Culture & Society*, 44(4):655–671.

Georgios Trichopoulos. 2023. Large language models for cultural heritage. In *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter*, pages 1–5.

Ted Underwood, Laura K Nelson, and Matthew Wilkens. 2025. Can language models represent the past without anachronism? *arXiv preprint arXiv:2505.00030*.

Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. 2025. Time blindness: Why video-language models can't see what humans can? *arXiv preprint arXiv:2505.24867*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yuqing Wang and Yun Zhao. 2023. Tram: Benchmarking temporal reasoning for large language models. *arXiv preprint arXiv:2310.00835*.

Zeqing Wang, Shiyuan Zhang, Chengpei Tang, and Keze Wang. 2025. Timecausality: Evaluating the causal ability in time dimension for vision language models. *arXiv preprint arXiv:2505.15435*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Joe Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 1963–1974.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix

A Task Specifications

600 multiple-choice questions in our benchmark is equally distributed across six temporal reasoning task categories (100 questions per category). Each task type evaluates different aspects of temporal awareness and cultural anachronism detection. Below we provide complete specifications for each task, including the number of images, answer structure, random baseline performance, and distractor construction methodology.

A.1 Chronological Sequencing

This task evaluates fundamental temporal ordering capabilities by requiring models to arrange four artifacts from different historical periods in correct chronological sequence from oldest to newest. Models must recognize visual cues indicating technological sophistication, artistic style evolution, and material availability across time periods.

Structure: 4 artifact images presented simultaneously. **Answer format:** One correct ordering from 24 possible permutations (4! = 24). **Random baseline:** 4.2% (1/24). **Distractor construction:** Artifacts are sampled from non-adjacent historical periods (minimum 2 periods apart) to ensure meaningful temporal separation and prevent trivial ordering based on obvious style differences.

A.2 Odd-One-Out Period Detection

This task assesses sensitivity to temporal anomalies by presenting four artifacts where three belong to one historical period and one belongs to a different period. Models must identify which artifact is temporally inconsistent with the others based on subtle visual cues in style, materials, manufacturing techniques, and artistic conventions.

Structure: 4 artifact images. **Answer format:** Single choice identifying the temporal outlier (1 of 4 options). **Random baseline:** 25.0% (1/4). **Distractor construction:** Three artifacts from the same historical period plus one artifact from a temporally adjacent period to increase difficulty and avoid obvious mismatches.

A.3 Period Grouping

This task examines pattern recognition across stylistic and material elements by requiring identification of three contemporaneous artifacts from a set of five. Models must recognize coherent aesthetic traditions, technological signatures, and cultural expressions characteristic of specific historical periods.

Structure: 5 artifact images. **Answer format:** Select 3 artifacts from the same period (10 possible combinations: $\binom{5}{3} = 10$). **Random baseline:** 10.0% (1/10). **Distractor construction:** Three artifacts from the target period plus two artifacts from different adjacent periods, requiring careful visual analysis to distinguish shared temporal characteristics.

A.4 Manufacturing Technique

This task tests understanding of historical production methods by asking which manufacturing techniques would have been available when a given artifact was created. Models must demonstrate knowledge of technological development timelines and craftsmanship evolution across Indian cultural history.

Structure: 1 artifact image with 4 technique options. **Answer format:** Multi-select (select all applicable techniques). **Scoring:** Exact-match required, all correct options must be selected and no incorrect options chosen. **Random baseline:** 6.25% (1/16 - one correct subset out of 16 possible selections). **Distractor construction:** Options include period-appropriate techniques, anachronistic modern techniques, and techniques from adjacent historical periods.

A.5 Material Availability

This task evaluates knowledge of historical material science by identifying which materials would NOT have been available during an artifact's creation period. Models must understand material discovery timelines, trade route development, and technological constraints across different eras.

Structure: 1 artifact image with 4 material options. **Answer format:** Multi-select identifying anachronistic materials (select all that would NOT have been available). **Scoring:** Exact-match required. **Random baseline:** 6.25% (1/16 - one correct subset out of 16 possible selections). **Distractor construction:** Options include period-appropriate materials, clearly anachronistic modern materials (e.g., plastic, acrylic for ancient artifacts), and materials from adjacent periods requiring nuanced knowledge of material availability timelines.

A.6 Style-Period Attribution

This task provides direct assessment of temporal classification capabilities by requiring models to

match artifacts with their correct historical periods. Models must recognize distinctive visual characteristics, artistic conventions, and cultural expressions associated with specific temporal periods in Indian cultural history.

Structure: 1 artifact image with 4 period options. **Answer format:** Single choice from 4 historical period options. **Random baseline:** 25.0% (1/4). **Distractor construction:** Four different historical periods spanning Indian history (e.g., Prehistoric, Modern India, Iron Age, Bronze Age), requiring fine-grained temporal discrimination.

Scoring methodology: All tasks employ exact-match accuracy as the primary metric. For single-choice tasks (Odd-One-Out, Style-Period Attribution), the model must select the single correct option. For multi-select tasks (Manufacturing Technique, Material Availability), the model must select ALL correct options and NO incorrect options, partial credit is not awarded. For Chronological Sequencing, the model must produce the exact correct ordering. For Period Grouping, the model must identify precisely the three contemporaneous artifacts.

B Prompt Templates

This section presents the standardized prompt templates used for evaluating vision-language models on temporal reasoning tasks with cultural heritage artifacts. Each template is designed to elicit specific aspects of historical understanding while maintaining consistency across all evaluated models. The prompts use `<image>` tags to indicate visual inputs, with the actual images provided during inference.

B.1 Chronological Sequence

The chronological sequencing task evaluates a model's fundamental ability to perceive and understand temporal progression through visual features alone. Models receive four artifacts from distinct historical periods and must arrange them from oldest to newest. This requires recognizing evolutionary patterns in artistic styles, technological sophistication, material usage, and iconographic conventions. The task is particularly challenging because it demands not just period identification, but understanding the relative temporal relationships between multiple artifacts simultaneously. Success requires the model to construct a coherent timeline based on visual evidence of technological development, stylistic maturation, and cultural evolution across Indian history.

Chronological Sequence Prompt Template

Below are 4 historical artifacts from different time periods.
Question: {question_text}
Please arrange these artifacts in chronological order from oldest to newest:
A. `<image>`
B. `<image>`
C. `<image>`
D. `<image>`
Provide the sequence from oldest to newest using letters (e.g., "A, C, B, D").

B.2 Odd One Out Period

This task assesses temporal anomaly detection by presenting four artifacts where three belong to the same historical period and one is from a different era. The challenge lies in identifying subtle visual cues that distinguish the outlier, whether through anachronistic stylistic elements, technological features, or material properties. Unlike simple classification tasks, this requires comparative analysis across multiple artifacts and recognition of shared temporal characteristics among the majority group. The task specifically tests whether models can detect deviations from expected period-consistent patterns, simulating real-world scenarios where cultural heritage experts must identify misattributed or misdated artifacts in museum collections.

Odd One Out Period Prompt Template

Below are 4 historical artifacts. One of them belongs to a different time period than the others.
Question: {question_text}
Which artifact belongs to a different time period than the others?
A. `<image>`
B. `<image>`
C. `<image>`
D. `<image>`
Answer with the option's letter (A, B, C, or D) from the given choices directly.

B.3 Material Availability

The material availability task evaluates understanding of technological timelines and resource avail-

Task	Images	Answer Format	Random Baseline
Chronological Sequence	4	Ordered sequence (24 permutations)	4.2%
Odd-One-Out Period	4	Single choice (1 of 4)	25.0%
Period Grouping	5	Select 3 of 5 (10 combinations)	10.0%
Manufacturing Technique	1	Multi-select from 4 options	6.25%
Material Availability	1	Multi-select from 4 options	6.25%
Style-Period Attribution	1	Single choice (1 of 4 periods)	25.0%

Table 3: Summary of task specifications. Chance baselines computed for single-answer and sequence tasks; multi-select tasks use exact-match scoring where all correct options must be selected.

ability across historical periods. Given a single artifact image, models must identify which materials from a provided list would NOT have been available during the artifact's creation period. This requires knowledge of when specific materials were discovered, developed, or became accessible in the Indian subcontinent. For instance, distinguishing whether an artifact could have been made with bronze (available from 3300 BCE) versus steel (developed much later) requires understanding metallurgical history. The task directly tests for material-based anachronisms, a common form of temporal misattribution where modern materials are incorrectly associated with ancient artifacts. The multi-select format allows for multiple correct answers, as several materials may be anachronistic for any given period.

Material Availability Prompt Template

Look at this historical artifact and answer the following question:
`<image>`
Question: {question_text}
Please select all options that apply from the following choices: {options_text}
List all correct letters separated by commas (e.g., "A, C" if options A and C are correct).

B.4 Manufacturing Technique

This task examines knowledge of historical production methods and technological capabilities. Models must determine which manufacturing techniques from a provided list would have been available when the artifact was created. The task requires understanding the evolution of craftsmanship methods, from primitive hand-molding and stone carving in prehistoric periods to sophisticated casting, glazing, and metalworking techniques in later eras. Success demands recognition of visual evidence indicating production methods, such as tool marks, construction joins, surface treatments, and structural features. The multi-select format reflects the reality that multiple techniques were often available simultaneously in different regions or for different purposes, requiring nuanced understanding of technological coexistence rather than simple chronological progression.

Manufacturing Technique Prompt Template

Look at this historical artifact and answer the following question:
`<image>`
Question: {question_text}
Please select all options that apply from the following choices: {options_text}
List all correct letters separated by commas (e.g., "A, C" if options A and C are correct).

B.5 Period Grouping

Period grouping tests pattern recognition across stylistic and material elements by requiring identification of three contemporaneous artifacts from a set of five. This task demands synthesis of multiple temporal cues, artistic conventions, technological signatures, material properties, and cultural motifs, to recognize coherent aesthetic traditions. Unlike the odd-one-out task which focuses on identifying a single outlier, this requires positive identification of shared temporal characteristics among multiple artifacts while recognizing that two artifacts belong to different periods. The task simulates museum curation scenarios where artifacts must be grouped

for period-specific exhibitions, requiring both individual dating expertise and comparative temporal analysis.

Period Grouping Prompt Template

Below are 5 historical artifacts. Three of them were likely created during the same time period.
Question: {question_text}
Which three artifacts were likely created during the same time period?
A. <image>
B. <image>
C. <image>
D. <image>
E. <image>
List the three letters corresponding to artifacts from the same period (e.g., "A, C, E").

B.6 Style-Period Attribution

Style-period attribution represents the most direct test of temporal classification capabilities. Given a single artifact, models must assign it to the correct historical period from four options spanning different eras of Indian history. This task requires recognition of distinctive visual characteristics, artistic conventions, iconographic traditions, and material signatures associated with specific temporal periods. The four-option format forces discrimination between periods that may share some overlapping features, requiring models to identify the most diagnostic temporal markers. Success on this task indicates mastery of period-specific aesthetic vocabularies and understanding of how artistic expressions evolved across different phases of Indian cultural development.

Style Period Attribution Prompt Template

Look at this historical artifact and answer the following question:
<image>
Question: {question_text}
Please select the correct option from the following choices: {options_text}
Answer with the option's letter (A, B, C, or D) from the given choices directly.

C Additional Analyses

This section provides supplementary analyses that complement the main results of Section 4. We examine (i) the factual-to-relational task spectrum, (ii) a manual failure-mode taxonomy over 96 incorrect GPT-4o responses, (iii) partial-credit evaluation, (iv) the reliability of our automatic evaluation pipeline, (v) the effects of temporal imbalance and textual metadata, and (vi) a cross-cultural pilot study on Western artifacts.

C.1 Factual vs. Relational Task Spectrum

The six tasks in TAB-VLM span a continuum from factual association (e.g., associating materials with historical periods) to relational temporal reasoning (e.g., ordering artifacts across periods). To quantify this continuum, we group the tasks into two categories: *fact-dominant* tasks (Material Availability, Style-Period Attribution, Manufacturing Technique), which can in principle be solved via memorized material–era mappings; and *temporal-dominant* tasks (Odd-One-Out Period, Period Grouping, Chronological Sequencing), which require reasoning over temporal relationships between multiple artifacts.

Task	Type	GPT-4o	Qwen2.5-VL-7B
Material Availability	Fact-Dom.	85.3	69.0
Style-Period Attribution	Fact-Dom.	60.3	58.8
Manufacturing Technique	Fact-Dom.	49.8	41.5
Odd-One-Out Period	Temp-Dom.	45.3	51.0
Period Grouping	Temp-Dom.	32.3	22.0
Chronological Sequencing	Temp-Dom.	30.7	14.0

Table 4: Accuracy (%) across the factual-to-relational task spectrum. Random baselines: Material Availability and Manufacturing Technique 6.25%; Style-Period Attribution and Odd-One-Out Period 25.0%; Period Grouping 10.0%; Chronological Sequencing 4.2%.

Performance drops sharply from fact-dominant to temporal-dominant tasks. GPT-4o exhibits a 54.6-point gap between Material Availability (85.3%) and Chronological Sequencing (30.7%), and this gradient is consistent across all ten models evaluated in Table 1. The gap is not attributable to missing cultural knowledge alone: smaller open-source models such as Qwen2.5-VL-3B score only 25.4% on Material Availability, indicating that the fact-dominant end is non-trivial at smaller scales. Conversely, even frontier models fail on temporal-dominant tasks (GPT-5.2: 37.2% on Chronological Sequencing). This pattern suggests that *relational temporal reasoning*, rather than *cultural*

knowledge retrieval, is the primary bottleneck in current VLMs on TAB-VLM.

C.2 Failure Mode Taxonomy

To characterise the nature of VLM errors on TAB-VLM, we manually analysed 96 incorrect GPT-4o responses, sampled uniformly at 16 per task. Each response was classified into one of three mutually exclusive failure types:

- **Knowledge gap**: incorrect or missing historical/cultural knowledge (e.g., wrong period attribution, incorrect material–era association).
- **Visual grounding failure**: misreading visual cues in the image (e.g., misidentifying stylistic features or material texture).
- **Relational temporal reasoning failure**: recognising individual artifacts correctly but failing at cross-artifact temporal comparison (e.g., describing each artifact's period correctly yet ordering them incorrectly).

Failure Type	Frequency
Knowledge Gap	29% (28/96)
Visual Grounding Failure	8% (8/96)
Relational Temporal Reasoning Failure	63% (60/96)

Table 5: Overall failure-type distribution across 96 incorrect GPT-4o responses.

Relational temporal reasoning failures dominate at 63%, followed by knowledge gaps at 29% and a small minority of visual grounding failures at 8%. The per-task breakdown (Table 6) shows that failure types are tightly aligned with task category.

Task	Category	Knowl.	Visual	Relational
Manufacturing Technique	Fact-Dom.	100%	0%	0%
Material Availability	Fact-Dom.	88%	12%	0%
Style-Period Attribution	Fact-Dom.	75%	19%	6%
Odd-One-Out Period	Temp-Dom.	25%	38%	37%
Period Grouping	Temp-Dom.	0%	6%	94%
Chronological Sequencing	Temp-Dom.	0%	0%	100%

Table 6: Per-task failure-type breakdown. Fact-dominant tasks fail almost exclusively via knowledge gaps; temporal-dominant tasks fail predominantly via relational reasoning errors.

Fact-dominant tasks produce knowledge-gap failures almost exclusively (75–100%), while temporal-dominant tasks produce relational-reasoning failures at 37–100%. Visual grounding failures are a small minority across all tasks. This task-aligned structure supports the interpretation in Section C.1: the two dominant failure modes are independent and separable, and poor performance on temporal-dominant tasks cannot be explained by missing cultural knowledge.

C.3 Partial-Credit Evaluation

The main results in Table 1 use exact-match accuracy, which assigns zero credit to partially correct answers (e.g., a chronological ordering that misplaces one of four artifacts). To test whether this strictness materially affects our conclusions, we re-scored the model outputs using partial-credit metrics appropriate to each task: Kendall's τ for Chronological Sequencing, multi-label F1 for the multi-select tasks (Material Availability, Manufacturing Technique), and Jaccard similarity for Period Grouping. Results are reported from a single run.

Model	Chron. Exact / τ	Mat. Avail. Exact / F1	Mfg. Tech. Exact / F1	Period Group. Exact / Jaccard
GPT-4o	30.4 / 0.620	85.6 / 0.945	48.9 / 0.810	32.3 / 0.564
Qwen2.5-VL-7B	14.0 / 0.407	69.0 / 0.935	37.6 / 0.750	21.0 / 0.495

Table 7: Exact-match vs. partial-credit metrics. Values in percent except τ and Jaccard, which are in $[0, 1]$.

Although partial-credit scores are higher in absolute terms, the relative ranking among models and the core conclusion (significantly deficient temporal reasoning on temporal-dominant tasks) remain unchanged. We retain exact-match as the primary metric in the main paper because cultural-heritage applications demand full correctness: a partially correct chronological ordering is still a misrepresentation of the historical record.

C.4 Reliability of the Automatic Evaluator

All results in Table 1 are produced by an automatic evaluation pipeline that parses model outputs and checks them against ground-truth answers. To assess the reliability of this pipeline, we conducted a human-evaluation study. Two authors of this paper independently annotated a random 10% sample of all model responses, labelling each response as either correctly or incorrectly scored by the pipeline. We then computed pairwise Cohen's Kappa (κ) between all three raters: the two human annotators and the automatic evaluator.

Human–human agreement ($\kappa = 0.92$) establishes an upper bound on inter-rater reliability for this task. The automatic evaluator achieves $\kappa = 0.85$ and $\kappa = 0.84$ against the two human annotators, approaching this ceiling. Under the standard interpretation guidelines of Landis and

Rater Pair	Cohen's κ
Human 1 $\leftrightarrow$ Human 2	0.92
Human 1 $\leftrightarrow$ Automatic Evaluator	0.85
Human 2 $\leftrightarrow$ Automatic Evaluator	0.84
Average Pairwise κ	0.87

Table 8: Pairwise Cohen's Kappa between two human annotators and the automatic evaluator.

Koch (Landis and Koch, 1977), all pairwise values fall in the "almost perfect" agreement range ($\kappa > 0.8$), indicating that the pipeline performs near human-level consistency in judging correctness.

C.5 Temporal Imbalance and Textual Metadata

The artifact counts in TAB-VLM vary across historical periods (Table 2), and our main evaluation uses visual input only. We examine both factors as potential confounds.

Temporal imbalance. We computed per-period accuracy on Style-Period Attribution for GPT-4o and Qwen2.5-VL-7B to test whether the number of artifacts per period predicts accuracy on that period.

Period	Count	GPT-4o	Qwen2.5-VL-7B
Classical (200 BCE – 650 AD)	21	76.9	57.1
Prehistoric (Before 3300 BCE)	21	75.0	66.7
Late Medieval (1200 – 1526 AD)	17	61.5	23.5
Bronze Age (3300 – 1800 BCE)	13	38.5	53.8
Iron Age (1800 – 200 BCE)	9	0.0	22.2
Early Modern (1526 – 1857 AD)	7	66.7	57.1
Modern India (1857 – Present)	6	66.7	66.7
Early Medieval (650 – 1200 AD)	6	0.0	33.3
Pearson r (count vs. acc.)	–	0.545 / 0.172	
p-value	–	0.162 / 0.684	

Table 9: Per-period accuracy (%) on Style-Period Attribution, and Pearson correlation between artifact count and accuracy. Correlations are weak and non-significant.

Correlations between artifact count and accuracy are weak and non-significant for both models ($p > 0.1$). The two hardest periods for GPT-4o (Iron Age and Early Medieval, both 0%) are mid-sized rather than the smallest, ruling out a frequency-driven explanation for the observed accuracy pattern.

Textual metadata. In realistic cultural-heritage deployments, artifact images are often accompanied by catalog metadata. To test whether our visual-only protocol understates model capability, we evaluated Qwen2.5-VL-7B on a 120-question sample (20 per task) with catalog metadata (object type and main material, excluding any period or dynasty information) appended to the visual prompt.

Model	Visual-Only	Visual+Metadata	Δ
Qwen2.5-VL-7B	40.1	40.8	+1.75%

Table 10: Effect of catalog metadata (object type and material) on accuracy (%). The gain is minimal, supporting the claim that the bottleneck is temporal reasoning rather than missing input context.

The gain of +1.75% is small relative to the performance gaps on temporal-dominant tasks. This is consistent with the interpretation in Section C.1 and Section C.2: the primary bottleneck is relational temporal reasoning rather than missing auxiliary input.

C.6 Cross-Cultural Pilot: Western Artifacts

The main benchmark focuses on Indian cultural heritage. To assess whether the observed patterns generalise beyond the Indian case and to provide preliminary evidence of a performance gap between Western and non-Western artifacts, we conducted a small pilot study using 84 Western artifacts (Greek/Roman, Medieval, Renaissance, Modern European) sourced from the Cleveland Museum of Art open-access collection. We evaluated GPT-4o on two representative tasks: Style-Period Attribution and Material Availability.

Artifact Source	Style-Period (%)	Material Avail. (%)
Indian (TAB-VLM)	60.3	85.3
Western (Pilot)	67.9	97.6
Δ	+7.6	+12.3

Table 11: GPT-4o accuracy on Style-Period Attribution and Material Availability for Indian artifacts (TAB-VLM) vs. a Western-artifact pilot (84 items, Cleveland Museum of Art). GPT-4o performs better on Western artifacts on both tasks.

GPT-4o scores +7.6 percentage points higher on Style-Period Attribution and +12.3 percentage points higher on Material Availability for Western artifacts relative to Indian artifacts. This gap is consistent with the hypothesis that non-Western cultural heritage presents additional challenges for current VLMs, likely reflecting the distribution of training data. We emphasise that this is a small pilot on two tasks and one model; a full cross-cultural extension of TAB-VLM to other traditions is left to future work.