MuseBench: A Comprehensive Benchmark for Multimodal Cultural Understanding of Chinese Museum Artifacts

Anonymous ACL submission

Abstract

Chinese museum artifacts represent a continuous cultural lineage and constitute a core component of global cultural heritage. Recent advances in Vision–Language models (VLMs) have shown promise in incorporating domain knowledge when applied to such collections. However, it remains unclear to what extent existing VLMs can effectively interpret and reason over professional museum artifact documentation. To address this gap, we introduce MuseBench, a comprehensive benchmark of Chinese museum artifacts, that evaluates two dimensions of VLM's cultural understanding: cultural reasoning and semantic alignments. MuseBench contains 128,592 images of 29,352 artifacts and 293,376 question-answer pairs, supporting two complementary tasks: Cultural Visual Question Answering and Cultural Retrieval. Through extensive evaluation of 25 mainstream VLMs, we observe that even the top-performing model achieves an average score below 22%, indicating substantial room for improvement. Our analysis shows significant performance variations across different tasks and identifies critical challenges primarily arising from professional terminology generation and structured metadata understanding. MuseBench thus provides a challenging benchmark with valuable insights that reveal substantial room for improvement in multimodal cultural understanding.

1 Introduction

Chinese museum artifacts represent a continuous cultural lineage. As emphasized by UNESCO, digitizing and interpreting these collections is critical for advancing scholarly research (UNESCO, 2015). Through digitization, museum artifacts form a vast repository of multimodal data, including images, textual descriptions, and structured metadata. In this context, recent advances in Vision-Language Models (VLMs) have shown promise in supporting cultural heritage tasks, such as cross-modal retrieval (Yuan et al., 2025; Zhang et al., 2025) and visual question answering (VQA) (Romero et al., 2024; Schneider et al., 2025). However, VLMs performance on professional museum artifact documentation remains largely underexplored.

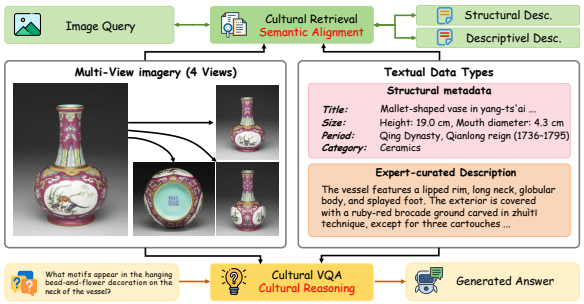

Figure 1: An example of the MuseBench. Each artifact includes standard multi-view imagery and two textual data types to support Cultural Retrieval and Cultural VQA tasks.

Despite their valuable resources, existing museum-related benchmarks remain insufficient for comprehensive evaluation of VLMs, as summarized in Table 1. Most benchmarks lack key data features required for professional museum analysis. For instance, VISCOUNTH (Becattini et al., 2023) incorporates official catalog sources and structured metadata, but does not provide multi-view imagery. Moreover, existing benchmarks typically support only a single evaluation task, focusing on either Cultural VQA or Cultural Retrieval, limiting unified assessment of cultural understanding. To fill this gap, we introduce **MuseBench**, a comprehensive professional multimodal benchmark for evaluating VLMs' cultural understanding of Chinese museum artifacts.

In the following, we explore MuseBench from both data and task perspectives. **(a) Data**: Different from existing benchmarks that rely primarily on single-source, web-derived data, MuseBench is curated from authoritative institutions (the Palace

Category	Feature	Museum-65 (Balauca et al., 2025)	VISCOUNTH (Becattini et al., 2023)	AQUA (Garcia et al., 2020)	EUFCC-CIR (Net and Gomez, 2024)	MuseBench (Ours)
Data	Official Catalog Source	✗	✓	✗	✗	✓
	Multi-View Imagery	✗	✗	✗	✗	✓
	Structured Metadata	✗	✓	✗	✗	✓
Task	Cultural VQA	✓	✓	✓	✗	✓
	Cultural Retrieval	✗	✗	✗	✓	✓

Table 1: Comparison of MuseBench with existing museum-oriented benchmarks.

Museum, the Zhejiang Provincial Museum, and the National Palace Museum), and constructed at a large scale, comprising 128,592 images of 29,352 artifacts and 293,376 QA pairs. As shown in Fig. 1, the benchmark further provides rich data representations by pairing standardized multi-view imagery, such as frontal, lateral, and detailed views, with expert-curated museum descriptions and structured metadata, including title, size, period, and category. **(b) Task**: MuseBench defines two complementary tasks: Cultural VQA for cultural reasoning and Cultural Retrieval for multimodal semantic alignments, to evaluate VLMs' cultural understanding. These tasks challenge models to establish cross-view associations across multiple visual perspectives. Subsequently, models are required to ground these visual cues within precise, domain-specific textual descriptions.

We conduct an extensive evaluation of 25 mainstream VLMs. The evaluated VLMs cover both closed-source systems such as GPT-5.1 (OpenAI, 2025) and Gemini-2.5 (Comanici et al., 2025), as well as open-source models including Qwen3-VL (Yang et al., 2025) and InternVL3.5 (Wang et al., 2025). Overall, current vision–language models exhibit limited cultural understanding in expert-curated museum settings. Current VLMs exhibit substantial limitations in cultural understanding within professional museum contexts. For the cultural VQA task, model-generated responses often fail to consistently produce the precise professional terminology required by the benchmark, even as model capacity increases. In the cultural retrieval task, embedding models tend to prioritize continuous, descriptive language representations over structured, label-oriented metadata, leading to degraded retrieval performance.

In summary, our contributions are as follows:

- We introduce MuseBench, a comprehensive benchmark on Chinese museum artifacts, covering 128,592 images of 29,352 artifacts with 293,376 QA pairs, curated from official museum institutions. The benchmark spans 47 artifact categories mapped to 18 dynasty-level periods.
- MuseBench paires standardized multi-view imagery with expert-curated museum descriptions and structured metadata to evaluate VLM's cultural understanding.
- We design comprehensive evaluation tasks, including Cultural VQA and Cultural Retrieval, to assess cultural reasoning and cross-modal semantic alignment.
- We conduct detailed analyses to reveal the key factors underlying the limited cultural understanding of VLMs across different tasks, offering valuable insights for future cultural heritage research.

2 Related Work

VLMs for Cultural Heritage. Vision-language models (VLMs) have advanced rapidly in recent years, including both closed-source systems such as GPT-5.1 (OpenAI, 2025) and Gemini-2.5 (Comanici et al., 2025), as well as large-scale open-source models such as Qwen3-VL (Yang et al., 2025) and InternVL3.5 (Wang et al., 2025). These models underpin a wide range of vision-language understanding tasks. In the cultural heritage domain, VLMs have been applied to tasks such as culturally grounded visual question answering (Romero et al., 2024), cross-modal retrieval (Zhang et al., 2025), and cultural caption generation for historical imagery (Ghaboura et al., 2025). Despite encouraging progress, it remains unclear how well these models perform on professionally curated museum data. This gap calls for museum-oriented benchmarks with professional artifact documentation.

Museum Multimodal Benchmarks. Existing museum-related benchmarks, such as MUSEUM-65 (Balauca et al., 2025) and VISCOUNTH (Be-

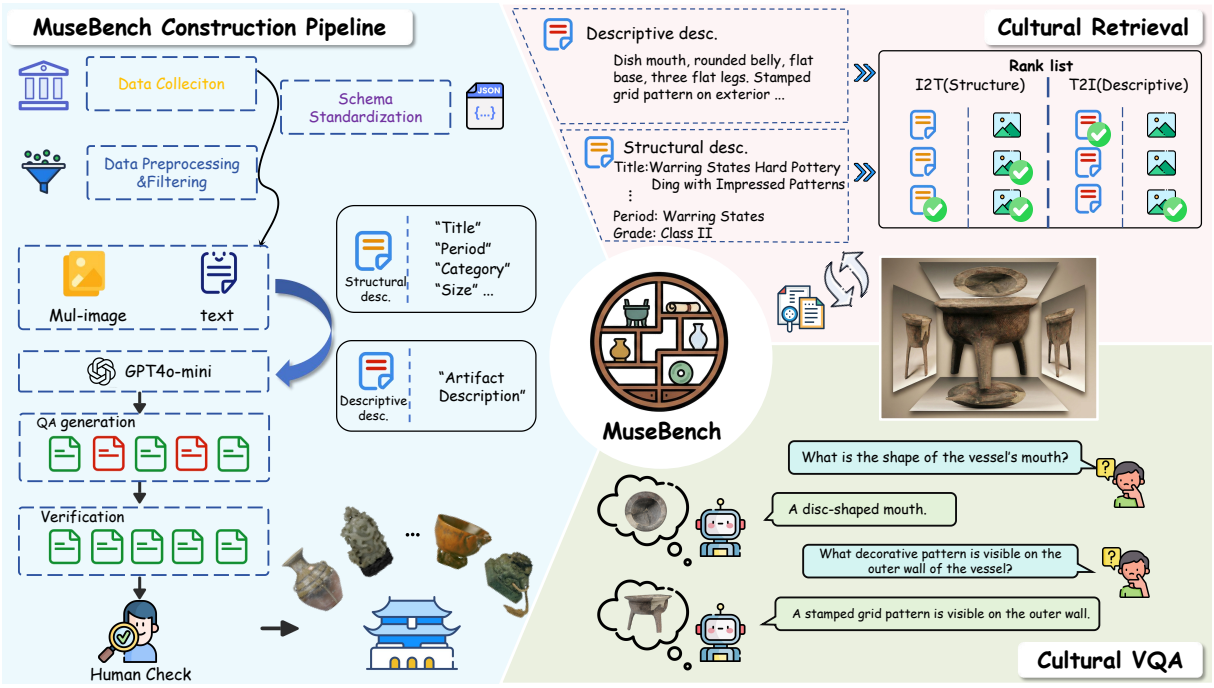

Figure 2: Overview of the MuseBench construction pipeline for two evaluation tasks, including Cultural VQA and Cultural retrieval.

cattini et al., 2023), leverage web-sourced data to evaluate broad multimodal coverage and cultural visual question answering. VISCOUNTH (Becattini et al., 2023) focuses on cultural visual question answering using image-text pairs collected from web sources. AQUA (Garcia et al., 2020) targets fine-art understanding by leveraging museum catalogs and art-historical annotations for supervision. More recent efforts, such as Seeing Cultural Heritage (Satar et al., 2025), further integrate structured metadata curated by domain experts to enhance annotation reliability. Despite their utility, current benchmarks primarily focus on general cultural context within limited task formats, potentially constraining the comprehensive evaluation of the latest VLMs.

3 MuseBench

We introduce MuseBench, a comprehensive multimodal benchmark for evaluating vision–language models on Chinese museum artifacts. MuseBench contains 128,592 images from 29,352 artifacts curated from authoritative museum collections, together with 293,376 QA pairs. Fig. 2 illustrates the MuseBench construction pipeline. Starting from museum data collection, artifact-level schema standardization, and data preprocessing and filtering, each artifact is organized into multi-view images, structured metadata, and expert-curated descriptions. Based on this unified representation, we construct two evaluation tasks: Cultural VQA, via question generation and verification, and Cultural Retrieval, under both structural and descriptive query settings.

3.1 Task Definition

Cultural VQA. Cultural VQA is designed to evaluate cultural reasoning over professionally curated museum artifacts. Given a multi-view image set of an artifact and a question, models are required to generate an answer grounded in observable visual evidence and consistent with authoritative museum descriptions. This task focuses on whether models can reason over fine-grained visual cues within culturally grounded museum contexts.

Cultural Retrieval. Cultural Retrieval in MuseBench evaluates cross-modal cultural semantic alignment in professional museum contexts. Models are required to associate professional museum documentation with multi-view visual representations. In practice, each individual image view serves as a query, while retrieval targets correspond to artifact-level multi-view image sets.

3.2 Data Construction

3.2.1 Data Collection and Standardization

Data Sources and Coverage. We collect artifact-level records from three major institutions: the

Zhejiang Provincial Museum, the Palace Museum, and the Taipei Palace Museum. Each record corresponds to a single museum artifact, and the initial corpus contains 618,939 images from 134,253 artifacts. The original collection covers 65 fine-grained artifact categories and spans Chinese history from the Neolithic to the People's Republic of China.

Artifact-Level Schema Standardization. To unify heterogeneous museum catalogs, we standardize all records into a fixed artifact-level JSON schema through three key steps. **(1) Artifact-level representation.** Each museum artifact is represented as a single record, aggregating all available visual views into a unified multi-view image set. This design ensures consistent alignment between visual evidence and textual descriptions at the artifact level. **(2) Dual-component textual schema.** Textual information is organized into two complementary components: *structural metadata* and *expert-curated descriptions*. Structural metadata includes canonical catalog fields such as title, category, period, size, author, and grade. Expert-curated descriptions consist of free-form curatorial texts written by museum professionals, describing observable visual characteristics and relevant cultural context. **(3) Category and period mapping.** To facilitate comparable analysis across museums, category and period labels are mapped to a unified taxonomy. Specifically, original artifact categories are unified into 47 normalized classes, and historical annotations are standardized into 18 dynasty-level periods, with detailed mappings provided in Appendix A.3 and A.4.

Data Preprocessing and Filtering. To ensure data quality, we apply preprocessing and filtering based on three criteria: **(1) Completeness filtering.** Artifacts missing two or more structural metadata fields or lacking an expert-curated description are removed. To support multi-view analysis, each retained artifact is required to contain at least two visual views. **(2) Temporal consistency checking.** The primary temporal annotation is the categorical field `period`. As a sanity check, any year-like token (YYYY) appearing in auxiliary text must not exceed the crawl year; records violating this constraint are discarded. **(3) Redundancy removal.** Near-duplicate image pairs within the same artifact are identified using two multimodal embedding models, tongyi-embedding-vision-plus and multimodal-embedding-v1. Candidate pairs with cosine similarity greater than 0.9 are treated as redundant and removed after expert verification. After preprocessing, each artifact retains at least two complementary visual views.

3.2.2 Cultural VQA Construction

Generation Method for VQA. To ensure both precision and semantic diversity, we combine rule-based templates with LLM-assisted generation for VQA construction across structured metadata and expert-curated descriptions. For structured metadata attributes, including `Title`, `Period`, `Category` and `Author`, we employ template-based generation. The `Size` field is explicitly excluded to avoid scale ambiguity. This approach yields high-precision QA pairs directly grounded in canonical museum records, resulting in a total of 117,408 structured-metadata-based questions.

For descriptive content, we generate questions from expert-curated descriptive briefs using GPT-4o-mini (Hurst et al., 2024). The generation prompt enforces three constraints: (1) *visual-centricity*, ensuring that questions require visual inspection rather than background knowledge alone; (2) *granularity*, focusing on isolated visual attributes instead of compound descriptions; and (3) *anonymization*, replacing artifact names with generic references to prevent data leakage. This process produces 208,399 descriptive QA pairs.

Two-Stage Quality Verification. All generated QA pairs undergo automated semantic filtering using GPT-4o (Hurst et al., 2024). The verifier is instructed to discard questions that can be answered without visual evidence or that are inconsistent with the associated images. After this automated stage, 293,376 out of 325,807 generated pairs are retained, corresponding to a pass rate of 90.04%, indicating that most questions are visually grounded and semantically consistent. To ensure data integrity, we perform human-in-the-loop verification via stratified sampling, manually inspecting 10% of samples across all categories. This manual verification stage achieves a pass rate of 98.16%, yielding 28,797 validated QA pairs, which constitute the golden set for Cultural VQA evaluation. Annotators evaluate visual dependency, image–answer consistency, and the presence of hallucinations; instances with identified issues are either corrected or discarded. Only validated pairs are included in the final benchmark.

3.2.3 Cultural Retrieval Construction

We design two complementary retrieval settings based on query formats: (1) Structural Retrieval,

which leverages metadata (e.g., title, period, and grade) to reflect the controlled information of professional catalogs; and (2) Descriptive Retrieval, which utilizes expert-curated natural language to capture visual characteristics and cultural context. For both settings, we evaluate bi-directional retrieval comprising Text-to-Image (T2I) and Image-to-Text (I2T) tasks to assess model capability in aligning artifact-level semantics with multi-view visual evidence under museum-level ambiguity.

3.3 Data Statistics

MuseBench contains 29,352 museum artifacts and 128,592 images, with each artifact documented by multiple complementary views, averaging 4.38 images per artifact. In total, the benchmark includes 293,376 QA pairs grounded in expert-curated museum descriptions and structured metadata. The number of QA pairs varies across institutions, reflecting differences in catalog richness and documentation practices. The artifact collection spans 47 standardized categories and covers major historical periods from the Neolithic era to the People's Republic of China, mapped into 18 dynasty-level periods, capturing the long-term continuity of Chinese material culture. Fig. 3 illustrates the dataset distribution across historical periods, museums, and major artifact categories.

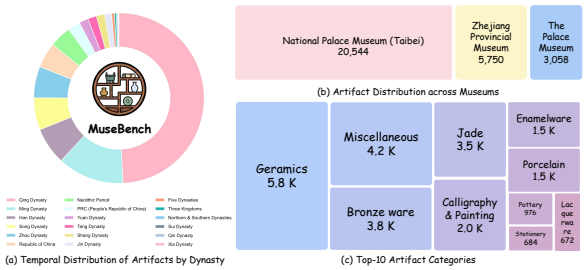

Figure 3: Statistics on MuseBench across three levels.

4 Experimental Details

4.1 Evaluated Methods

We evaluate 25 mainstream VLMs on MuseBench, comprising 17 models for Cultural VQA and 8 dedicated to Cultural Retrieval. These include: **(a) Closed-source models** such as Gemini-2.5-Pro and Gemini-2.5-Flash (Comanici et al., 2025), GPT-5.1 (OpenAI, 2025), and Doubao-seed-1.6-thinking (Seed, 2025). **(b) Open-source models**, including Qwen2.5-VL series (Bai et al., 2025), Qwen3-VL series (Yang et al., 2025), InternVL3 series (Zhu et al., 2025) and InternVL3.5 series (Zhu et al., 2025; Wang et al., 2025), as well as DeepSeek-VL2-Tiny (Wu et al., 2024). **(c) Embedding models** such as CN-CLIP (Yang et al., 2022), multimodal-embedding-v1, tongyi-embedding-vision-flash, tongyi-embedding-vision-plus, and the Qwen2.5-VL embedding series (Bai et al., 2025).

4.2 Evaluation Metrics

For Cultural VQA, **ACC (Accuracy)** is used for questions targeting structured attributes with a single correct answer, such as dynasty period or author. **BERT (BERTScore)** measures semantic similarity between model-generated answers and expert-authored references, capturing meaning-level alignment beyond exact wording. **ANLS** measures string-level similarity between predicted and reference answers. For Cultural Retrieval, we report **Recall@K**, which measures the proportion of queries for which the ground-truth artifact appears among the top-K retrieved candidates.

4.3 Experiment Setup

All experiments are conducted under a zero-shot setting without task-specific fine-tuning. For open-source models with parameter sizes below 9B, as well as CN-CLIP, are conducted on two NVIDIA A6000 GPUs. Other open-source and all closed-source models are evaluated via official APIs using default inference settings.

5 Results and Analysis

5.1 Main Results

Overall performance leaves substantial room for improvement. For Cultural VQA, the average score across all 17 models is merely 32.53%, underscoring the significant difficulty of the task. Except for the top-performing Qwen3VL-235B-A22B-Instruct (36.89%), all other evaluated models fall below 35%, as shown in Table 2. The performance limitation suggests that most models still struggle with the complex reasoning required to generate professionally grounded answers aligned with museum catalogs. Regarding Cultural Retrieval (Table 4), the top-performing Qwen2.5-VL embedding model, utilizing 2048-dimensional representations, achieves a mean R@1 of only 6.7% across all settings. This underscores a significant gap in aligning artifact-level semantics with visual evidence.

Model	Overall			Zhejiang Provincial Museum			The Palace Museum			National Palace Museum		
	Score	BERT	ANLS	Score	BERT	ANLS	Score	BERT	ANLS	Score	BERT	ANLS
Closed-source Models												
Gemini-2.5-Pro (Comanici et al., 2025)	34.96	67.41	4.25	36.56	67.47	5.78	35.27	68.16	7.05	34.52	67.25	3.33
Gemini-2.5-Flash (Comanici et al., 2025)	34.34	66.37	3.98	35.48	66.35	5.61	34.15	66.54	5.94	34.11	66.34	3.20
GPT-5.1 (OpenAI, 2025)	34.47	67.16	2.75	34.09	66.47	3.72	34.00	67.25	4.04	34.66	67.30	2.27
Doubao-seed-1.6-Thinking (Seed, 2025)	34.49	67.49	5.36	36.50	67.42	7.98	35.19	67.61	7.04	33.87	67.49	4.40
Open-source Models												
Qwen2.5VL-3B-Instruct (Bai et al., 2025)	30.83	62.06	2.57	35.77	63.22	5.62	31.67	63.34	4.54	29.49	61.53	1.45
Qwen2.5VL-7B-Instruct (Bai et al., 2025)	32.08	64.94	2.60	34.82	65.30	5.14	31.81	64.67	3.13	31.48	64.91	1.89
Qwen2.5VL-72B-Instruct (Bai et al., 2025)	33.46	66.51	4.82	36.01	66.96	7.41	32.94	66.02	5.57	32.96	66.50	4.06
Qwen3VL-8B-Instruct (Yang et al., 2025)	33.36	67.20	4.19	36.67	67.94	8.07	33.76	66.92	5.52	32.50	67.08	3.01
Qwen3VL-30B-A3B-Instruct (Yang et al., 2025)	34.57	67.86	5.02	38.05	68.61	8.39	34.63	67.82	5.70	33.73	67.69	4.09
Qwen3VL-32B-Instruct (Yang et al., 2025)	34.60	68.02	5.19	37.77	68.59	8.04	35.03	67.97	6.68	33.76	67.89	4.21
Qwen3VL-235B-A22B-Instruct (Yang et al., 2025)	**36.89**	**68.98**	**7.44**	**39.93**	**69.74**	**11.64**	**37.75**	**69.06**	**9.73**	**36.00**	**68.78**	**5.99**
Qwen3VL-235B-A22B-Thinking (Yang et al., 2025)	34.97	67.43	5.53	37.74	68.23	8.93	35.09	67.54	7.74	34.28	67.21	4.29
InternVL3-2B (Zhu et al., 2025)	27.76	58.39	0.25	29.16	58.96	0.77	28.92	60.29	0.60	27.20	57.88	0.05
InternVL3-8B (Zhu et al., 2025)	31.23	64.27	2.73	31.82	63.50	3.60	31.95	64.92	3.48	30.95	64.32	2.38
InternVL3.5-1B (Wang et al., 2025)	27.63	58.91	0.51	28.95	59.16	1.45	28.71	59.78	0.93	27.10	58.68	0.21
InternVL3.5-2B (Wang et al., 2025)	28.45	60.41	0.96	29.81	61.70	2.07	29.77	61.51	1.98	27.86	59.89	0.49
Deepseek-VL-2-Tiny (Wu et al., 2024)	28.89	61.20	0.76	29.44	60.28	1.02	29.79	60.60	1.15	28.58	61.54	0.62

Table 2: Cultural VQA results on the MuseBench golden set (%). **Overall** is computed as a QA-count-weighted average over the three museum subsets. Best results are shown in **bold**.

Model	BERT (%)	ANLS (%)	Score (%)
Closed-source			
Gemini-2.5-Pro	67.64	3.98	35.81
Gemini-2.5-Flash	66.64	3.54	35.09
GPT-5.1	67.03	2.38	34.71
Doubao-seed-1.6	67.49	4.83	36.16
Open-source			
Qwen2.5VL-7B-Instruct	64.90	2.57	33.74
Qwen2.5VL-72B-Instruct	66.39	4.87	35.63
Qwen3VL-235B-A22B-Instruct	**69.12**	**8.88**	**38.99**
Qwen3VL-235B-A22B-Thinking	67.62	6.63	37.12
InternVL3-8B	64.01	3.09	33.55
InternVL3.5-2B	60.63	1.79	31.21
DeepSeek-VL2-Tiny	61.88	1.02	31.45

Table 3: Overall performance on descriptive Cultural VQA evaluation over the MuseBench golden set. Results are computed as QA-count-weighted averages across Taiwan (15,882), the Palace Museum (3,085), and Zhejiang Provincial Museum (3,298).

Structured metadata pose significant challenge to VLMs' cultural understanding. In Table 2, BERTScore for Qwen3VL-235B-A22B-Instruct exceeds its ANLS score by nearly ninefold, as the former prioritizes semantic similarity while the latter requires exact string matching. This discrepancy indicates that while models capture the general semantic intent of queries, they remain inadequate in producing the catalog-level terminology required by the benchmark. In addition, structural information poses greater challenges for model understanding than descriptive information. To assess the impact of structured metadata, we evaluate VLMs performance by removing all structured metadata from the VQA setting. Across all models, this removal consistently improves performance by an average of 1.57%, as shown in Table 3. For the best-performing model, Qwen3VL-235B-A22B-Instruct, BERTScore and ANLS increase by 0.14% and 1.44%, respectively, compared with the overall VQA setting. Furthermore, we reformulate structured metadata into natural language queries for retrieval embedding, which yields consistent performance gains across models, improving R@1 by an average of 6.6%, as shown in Fig. 4.

Scaling and reasoning provide limited improvements to VLMs cultural understanding. Although larger models consistently outperform smaller counterparts across all museum subsets, for example, Qwen3VL-235B-A22B-Instruct exceeds Qwen3VL-8B-Instruct by over 3.5% in overall score, even the largest models remain far below expert level, with scores below 40%. This indicates that model scaling alone is insufficient to close the gap. Moreover, comparisons between Instruct and Thinking variants show that explicit reasoning does not yield systematic improvements. Notably, Qwen3VL-235B-A22B-Thinking underperforms its Instruct counterpart by nearly 2%. This suggests that Cultural VQA relies more on fine-grained visual grounding and domain knowledge than on extended reasoning traces.

5.2 Further Analysis in Cultural VQA

Closed-source models are competitive but lack domain-specific superiority. Closed-source systems such as Gemini-2.5 (Comanici et al., 2025) and GPT-5.1 (OpenAI, 2025) achieve Overall

Model	Text dim	Image dim	Structural Retrieval						Descriptive Retrieval					
			I2T			T2I			I2T			T2I		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CN-CLIP	1024	1024	3.31	9.85	14.86	3.91	10.50	15.50	3.01	8.93	13.34	4.25	10.56	15.71
multimodal-embedding-v1	1024	1024	**6.00**	**15.86**	**22.44**	**6.31**	**13.74**	18.84	5.25	13.13	18.15	7.93	15.78	21.00
tongyi-embedding-vision-flash	768	768	1.59	4.97	7.67	1.09	3.27	5.18	2.63	7.26	10.64	2.50	6.07	8.58
tongyi-embedding-vision-plus	1152	1152	2.68	7.88	11.68	2.00	5.51	8.36	4.21	10.77	15.33	4.10	9.54	13.44
qwen2.5-vl-embedding	512	512	4.76	13.59	19.68	4.33	10.78	15.38	5.38	14.21	20.15	8.00	16.87	22.23
qwen2.5-vl-embedding	768	768	5.04	14.03	20.10	5.00	11.95	16.78	5.73	14.77	20.76	8.62	17.98	23.78
qwen2.5-vl-embedding	1024	1024	5.24	14.53	20.73	5.06	12.26	17.30	5.90	**15.22**	21.39	9.09	18.55	24.49
qwen2.5-vl-embedding	2048	2048	5.48	15.04	21.35	5.88	13.68	**19.31**	**5.95**	15.17	**21.45**	**9.55**	**19.51**	**25.62**

Table 4: Cultural retrieval performance on MuseBench in terms of R@K (%). Results are reported for Structural and Descriptive queries under I2T and T2I settings. Text dim and Image dim denote embedding dimensionalities. Best results are in **bold**.

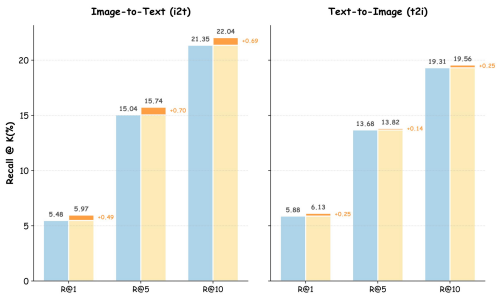

Figure 4: Comparison of cultural retrieval performance using structural metadata (blue) and their natural-language reformulations (yellow). Orange bars indicate the absolute performance gains.

Model	Category BERT (%)	Period Acc (%)	Author Acc (%)
Closed-source			
Gemini-2.5-Pro	73.13	24.70	15.01
Gemini-2.5-Flash	72.78	24.54	9.36
GPT-5.1	71.83	**27.62**	1.80
Doubao-seed-1.6-Thinking	71.52	15.79	14.50
Open-source			
Qwen2.5VL-7B-Instruct	70.92	15.67	14.75
Qwen2.5VL-72B-Instruct	**74.18**	12.71	16.50
Qwen3VL-235B-A22B-Instruct	71.64	22.53	**25.47**
Qwen3VL-235B-A22B-Thinking	71.37	19.19	25.06
InternVL3-8B	69.94	8.73	5.97
InternVL3.5-2B	71.99	7.11	1.82
DeepSeek-VL2-Tiny	72.43	7.43	5.78

Table 5: Cultural VQA performance comparison across category, period, and author prediction tasks.

Scores comparable to those of large open-source models. However, none of the evaluated closed-source models outperform Qwen3VL-235B-A22B-Instruct (Yang et al., 2025). This indicates that proprietary resources fail to yield a decisive advantage here, potentially due to gaps in specialized data and cultural linguistic alignment

Models exhibit limited capability in inferring deep cultural information from structured metadata. We evaluate VLMs' cultural understanding across category recognition and deeper author and period prediction for paintings and calligraphy. As shown in Table , while category recognition achieves substantially higher performance, with BERTScore consistently exceeding 70%, author and period prediction remain markedly weaker, with best accuracies below 30%. This indicates that although models perform well on high-level category recognition, they struggle to infer fine-grained cultural attributes from visually subtle cues.

5.3 Further Analysis in Cultural Retrieval

VLMs semantic alignment remains critically insufficient for complex cultural artifacts across all retrieval tasks. As illustrated in Table 4, multimodal-embedding-v1 yields the highest performance for Structural Retrieval, with R@1 scores of only 6.31% and 6% in T2I and I2T settings, respectively. This gap highlights the difficulty of aligning multi-view images with structured metadata. Conversely, Descriptive Retrieval relies more on semantically rich representations, where the Qwen2.5-VL 2048-dimensional variant achieves the peak R@1 of 9.55%. Overall, these results highlight a fundamental limitation in cross-modal semantic alignment for complex cultural artifacts.

Fine-tuning significantly mitigates the alignment gap in cultural retrieval. In the zero-shot setting, all models perform poorly on both Structural and Descriptive Retrieval. As shown in Table 4, the zero-shot R@1 of CN-CLIP remains around 3~4% across both image-to-text (I2T) and text-to-image (T2I) settings, with its R@10 staying below 16%. This indicates that off-the-shelf multimodal embeddings struggle to retrieve culturally relevant artifacts when directly applied to museum data. As illustrated in Fig. 5, fine-tuning yields substantial improvements. For Structural Retrieval, CN-

CLIP's Recall@1 surges from 3.31% to 24.61% (I2T) and 24.52% (T2I), with R@10 improving by over 50 percentage points. A similar trajectory is observed for Descriptive Retrieval, where the model's Recall@1 rises to over 22%, demonstrating the necessity of domain-specific adaptation.

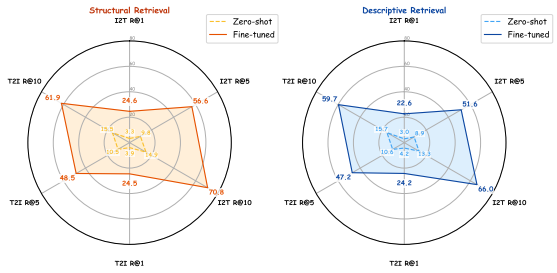

Figure 5: Comparison of cultural retrieval performance on CN-CLIP under zero-shot and fine-tuned settings.

5.4 Case Study

Fig. 6 illustrates a multi-view artifact example along side two representative Cultural VQA cases on MuseBench.

Case 1 Structured metadata: period prediction with multi-view inputs. The ground-truth label is the Qing dynasty *Qing dynasty* (清代), and the model predicts the *Yongzheng reign of the Qing dynasty* (清·雍正), which is considered correct under dynasty-level evaluation. Crucially, this prediction is supported by evidence visible in the bottom view of the artifact, where an inscribed reign mark is clearly present. Such information is not observable from the frontal or side views, which mainly convey decorative style rather than explicit temporal markers. When this bottom view is removed, the remaining views no longer provide sufficient cues for identifying the specific reign, and the model fails to produce a correct period prediction. This example indicates that VLMs benefit from multi-view visual cues but remain reliant on explicit markers for historical reasoning.

Case 2 Descriptive understanding: identifying the artifact's exterior decorative technique. The expert reference explicitly specifies *falangcai enamels on a red ground* (红地珐琅彩), together with distinctive catalog terms such as *gilt-outlined medallions* (描金边饰) and *cartouches with auspicious motifs* (开光内吉祥纹样). The Instruct model captures the coarse meaning (enamel on a red background) and therefore obtains a relatively high BERTScore. However, it omits critical museum-specific terminology (e.g., 珐琅彩/开光/描金), leading to an ANLS of 0 due to low string-level overlap. The Thinking variant introduces plausible but unverifiable additions (e.g., "imperial kilns") and further drifts from the reference phrasing, which reduces BERTScore and does not improve ANLS. This example illustrates that, in museum contexts, models may appear semantically aligned while still failing to produce terminology-precise descriptions required by expert documentation.

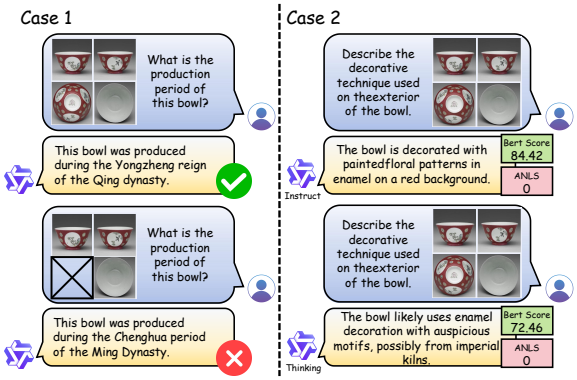

Figure 6: Case study of Cultural VQA involving multi-view inputs.

6 Conclusion

This paper introduces MuseBench, a comprehensive multimodal benchmark specifically designed for comprehensively evaluating VLMs for Chinese museum artifacts. It covers artifacts from three official museum catalogs spanning 47 categories and 18 dynasty-level periods, with two complementary tasks for comprehensive evaluation of VLMs' cultural understanding in complex cultural contexts. Experimental evaluations on 25 mainstream models indicate substantial room for improvement in museum-level cultural understanding. In addition, we provide an in-depth analysis of experimental results across tasks and dissect the underlying influencing factors, offering actionable insights for future research. We expect MuseBench to serve as a challenging benchmark for evaluating VLMs and fostering progress in cultural interpretation and reasoning for digital cultural heritage.

Limitations

MuseBench evaluates Cultural Visual Question Answering using automatic metrics aligned with expert-authored museum catalog annotations. While suitable for large-scale and reproducible evaluation, descriptions from different museums

may exhibit minor variations for similar cultural concepts.

Ethical Consideration

MuseBench is constructed based on publicly accessible museum catalogs and is intended solely for academic research on multimodal understanding of cultural heritage. When using this benchmark, researchers should be aware that museum descriptions reflect institutional curatorial perspectives, which may vary across museums and historical contexts. Care should be taken to avoid cultural misinterpretation, oversimplification, or the reinforcement of biased or anachronistic views when analyzing or deploying model outputs. All data sources used in MuseBench comply with copyright and usage policies of the originating institutions. The dataset must not be used for commercial purposes or applications that conflict with the principles of cultural respect, scholarly integrity, or responsible use of artificial intelligence.

References

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

Ada-Astrid Balauca, Sanjana Garai, Stefan Balauca, Rasesh Udayakumar Shetty, Naitik Agrawal, Dhwanil Subhashbhai Shah, Yuqian Fu, Xi Wang, Kristina Toutanova, Danda Pani Paudel, and 1 others. 2025. Understanding museum exhibits using vision-language reasoning. In *ICCV*, pages 2227–2238.

Federico Becattini, Pietro Bongini, Luana Bulla, Alberto Del Bimbo, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. 2023. Viscounth: a large-scale multilingual visual question answering dataset for cultural heritage. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–20.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. 2020. A dataset and baselines for visual question answering on art. In *ECCV*, pages 92–108. Springer.

Sara Ghaboura, Ketan Pravin More, Ritesh Thawkar, Wafa Al Ghallabi, 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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23627–23641.

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*.

Francesc Net and Lluis Gomez. 2024. Eufcc-cir: A composed image retrieval dataset for glam collections. In *ECCV*, pages 196–211. Springer.

OpenAI. 2025. Gpt-5 system card. https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf. System-level documentation for the GPT-5 model.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *NeurIPS*, 37.

Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. Seeing culture: A benchmark for visual reasoning and grounding. In *EMNLP*, pages 22238–22254.

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. GIMMICK: Globally inclusive multimodal multitask cultural knowledge benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9605–9668.

ByteDance Seed. 2025. Seed1. 6 tech introduction. *Accessed on September*, 28:2025.

UNESCO. 2015. Recommendation concerning the protection and promotion of museums and collections, their diversity and their role in society. Adopted by the General Conference of UNESCO, Paris, France. 17 November 2015.

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

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Junyi Yuan, Jian Zhang, Fangyu Wu, Huanda Lu, Dongming Lu, and Qiufeng Wang. 2025. Towards cross-modal retrieval in chinese cultural heritage documents: Dataset and solution. In *ICDAR*, pages 570–586. Springer.

Jian Zhang, Junyi Guo, Junyi Yuan, Huanda Lu, Yanlin Zhou, Fangyu Wu, Qiufeng Wang, and Dongming Lu. 2025. Llm-driven completeness and consistency evaluation for cultural heritage data augmentation in cross-modal retrieval. In *EMNLP*, pages 19418–19428.

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

A: Source Catalog Formats and Standardized Schema

A.1 Original Catalog Formats

Listing 1: Original catalog format - Zhejiang Provincial Museum (English translation).

```
{
    "images": [
        "String (URL)",
        "String (URL)",
        "..."
    ],
    "title": "String",
    "category": "String",
    "period": "String",
    "size_cm": "String",
    "description": "String",
    "grade": "String"
}
```

Listing 2: Original catalog format — Zhejiang Provincial Museum (Chinese version).

```
{
    "图片": [
        "String (URL)",
        "String (URL)",
        "..."
    ],
    "名称": "String",
    "馆藏类型": "String",
    "藏品年代": "String",
    "尺寸(cm)": "String",
    "简介": "String",
    "藏品级别": "String"
}
```

Listing 3: Original catalog format - The Palace Museum (English translation).

```
{
    "category": "String",
    "title": "String",
    "author": "String",
    "period": "String",
    "description": "String",
    "images": [
        "String (URL)",
        "String (URL)",
        "..."
    ]
}
```

Listing 4: Original catalog format — The Palace Museum (Chinese version).

```
{
    "类型": "String",
    "文物名称": "String",
    "作者": "String",
    "年代": "String",
    "简介": "String",
    "图片": [
        "String (URL)",
        "String (URL)",
        "..."
    ]
}
```

Listing 5: Original catalog format - National Palace Museum (English translation).

```
{
    "images": [
        "String (URL)",
        "String (URL)",
        "..."
    ],
    "unified artifact number": "String",
    "title": "String",
    "category": "String",
    "period": "String",
    "size": "String",
    "description": "String"
}
```

Listing 6: Original catalog format - National Palace Museum (Chinese version).

```
{
    "片": [
        "String (URL)",
        "String (URL)",
        "..."
    ],
    "文物一": "String",
    "品名": "String",
    "分": "String",
    "代": "String",
    "尺寸": "String",
    "明": "String"
}
```

A.2 Standardized Object-Level JSON Schema

Listing 7: Standardized object-level JSON schema — Zhejiang Provincial Museum (English translation).

```
{
  "id": "Integer",
  "structural_desc.": {
    "title": "String",
    "category": "String",
    "period": "String",
    "size_cm": "String",
    "grade": "String"
  },
  "descriptive_desc.": {
    "description": "String"
  },
  "images": [
    "String (URL)",
    "String (URL)",
    "..."
  ]
}
```

Listing 8: Standardized object-level JSON schema — Zhejiang Provincial Museum (Chinese version).

```
{
  "id": "Integer",
```

```
  "structural_desc.": {
    "名称": "String",
    "藏馆类型": "String",
    "藏品年代": "String",
    "尺寸（cm）": "String",
    "藏品级别": "String"
  },
  "descriptive_desc.": {
    "简介": "String"
  },
  "images": [
    "String (URL)",
    "String (URL)",
    "..."
  ]
}
```

Listing 9: Standardized object-level JSON schema — The Palace Museum (English translation).

```
{
  "id": "Integer",
  "structural_desc.": {
    "title": "String",
    "category": "String",
    "period": "String",
    "author": "String",
  },
  "descriptive_desc.": {
    "description": "String"
  },
  "images": [
    "String (URL)",
    "String (URL)",
    "..."
  ]
}
```

Listing 10: Standardized object-level JSON schema — The Palace Museum (Chinese version).

```
{
  "id": "Integer",
  "structural_desc.": {
    "文物名称": "String",
    "类型": "String",
    "时代": "String",
    "作者": "String",
  },
  "descriptive_desc.": {
    "简介": "String"
  },
  "images": [
    "String (URL)",
    "String (URL)",
    "..."
  ]
}
```

Listing 11: Standardized object-level JSON schema — National Palace Museum (English translation).

```
{
  "id": "Integer",
  "unified artifact number": "String",
  "structural_desc.": {
    "title": "String",
    "category": "String",
    "period": "String",
    "size": "String",
  },
  "descriptive_desc.": {
    "description": "String"
  },
  "images": [
    "String (URL)",
    "String (URL)",
    "..."
  ]
}
```

Listing 12: Standardized object-level JSON schema — National Palace Museum (Chinese version).

```
{
  "id": "Integer",
  "文物一": "String",
  "structural_desc.": {
    "品名": "String",
    "分": "String",
    "代": "String",
    "尺寸": "String",
  },
  "descriptive_desc.": {
    "明": "String"
  },
  "片": [
    "String (URL)",
    "String (URL)",
    "..."
  ]
}
```

A.3 Dynasty-Level Period Mapping and Statistics

Table 6 outlines the mapping between dynasties and their corresponding time periods used in MuseBench. The second column summarizes the distribution of artifacts across dynasties for the entire dataset, while the third, fourth, and fifth columns present detailed chronological statistics for the three individual sub-datasets: the Zhejiang Provincial Museum, the Palace Museum, and the National Palace Museum.

Our statistics reveal a significant long-tailed distribution in the chronological data of museum artifacts. Specifically, the majority of samples are concentrated in recent dynasties, such as the Qing and Ming. In contrast, ancient or short-lived dynasties such as the Xia, Qin, and Sui form the "tail" with extremely scarce samples. This imbalance objectively reflects the historical reality that fewer artifacts survive from older periods, as well as the specific collection focuses of different museums.

Dynasty	Year Range	MuseBench	Museums (Artifact Count)		
			Zhejiang Prov. Museum	Palace Museum	National Palace Museum
Neolithic Period	–	1,197	540	27	630
Xia Dynasty	2070–1600 BCE	17	11	2	4
Shang Dynasty	1600–1046 BCE	469	40	29	400
Zhou Dynasty	1046–221 BCE	1,738	270	129	1,339
Qin Dynasty	221–207 BCE	25	1	8	16
Han Dynasty	206 BCE–220 CE	2,068	206	134	1,728
Three Kingdoms	220–265 CE	97	25	7	65
Jin Dynasty	265–420 CE	381	322	21	38
Northern and Southern Dynasties	420–589 CE	75	29	19	27
Sui Dynasty	581–618 CE	34	9	4	21
Tang Dynasty	618–907 CE	480	155	61	264
Five Dynasties	907–960 CE	158	118	13	27
Song Dynasty	960–1279 CE	1,794	498	254	1,042
Yuan Dynasty	1271–1368 CE	512	128	55	329
Ming Dynasty	1368–1644 CE	3,703	347	545	2,811
Qing Dynasty	1644–1911 CE	14,420	1,337	1,678	11,405
Republic of China	1912–1949 CE	1,414	1,115	4	295
People's Republic of China	1949–present	770	599	68	103
Total		**29,352**	**5,750**	**3,058**	**20,544**

Table 6: Dynasty-level period mapping and artifact counts across MuseBench and three museums.

A.4 Category Mapping and Statistics

Museum catalogs from different institutions use heterogeneous category systems with different naming conventions and levels of granularity. To support consistent dataset analysis and cross-museum comparison, we map these museum-specific categories to a unified taxonomy for statistical reporting only. Importantly, this mapping does not modify the original annotations: all Cultural Retrieval and Cultural Visual Question Answering evaluations are conducted using the original category labels provided by each museum.

Original	Gugong	Taipei	Zhejiang	Total	Mapping
陶瓷器	0	5779	0	5779	陶瓷器
杂项	0	4184	0	4184	杂项、其他
铜器	0	3812	0	3812	铜器
玉器	0	3541	0	3541	玉石器、宝石、玉器
书法、绘画	0	0	1998	1998	书法、绘画
珐琅器	0	1513	0	1513	珐琅器、珐琅
瓷器	0	0	1468	1468	瓷器
陶瓷	976	0	0	976	陶瓷器
文具	0	600	84	684	文具、文房用品
漆器	112	478	82	672	漆器
雕刻	0	462	0	462	雕刻、雕塑、造像
玺印	300	0	0	300	玺印、玺印符牌
碑帖拓本	0	0	286	286	碑帖拓本
织绣	242	0	15	257	织绣
玉石器、宝石	0	0	246	246	玉石器、宝石、玉器
玺印符牌	0	0	244	244	玺印、玺印符牌
铭刻	243	0	0	243	铭刻
牙骨角器	0	0	198	198	牙骨角器
竹木牙角匏	181	0	0	181	竹木牙角匏
陶器	0	0	178	178	陶器
铜器	0	0	177	177	铜器

(To be continued...)

(Continued from previous column)

Original	Gugong	Taipei	Zhejiang	Total	Mapping
钱币	0	0	163	163	钱币
青铜器	140	0	0	140	青铜器
石器、石刻、砖瓦	0	0	134	134	石器、石刻、砖瓦
玉石器	131	0	0	131	玉石器、宝石、玉器
文件、宣传品	0	0	107	107	文件、宣传品
宫廷宗教	102	0	0	102	宫廷宗教
文房用品	101	0	0	101	文具、文房用品
雕塑、造像	0	0	89	89	雕刻、雕塑、造像
绘画	81	0	0	81	书法、绘画
钟表仪器	81	0	0	81	钟表仪器
织品	0	77	0	77	织品
档案文书	0	0	74	74	档案文书
家具	54	0	16	70	家具
珐琅	66	0	0	66	珐琅器、珐琅
法书	0	56	0	56	书法、绘画
生活器具	50	0	0	50	生活器具
金银器	0	0	47	47	金银锡器
音乐戏曲	44	0	0	44	音乐戏曲
书法	40	0	0	40	书法、绘画
雕塑	32	0	0	32	雕刻、雕塑、造像
武器	0	0	31	31	武器
竹木雕	0	0	29	29	竹木雕
武备仪仗	28	0	0	28	武备仪仗
标本、化石	0	0	24	24	标本、化石
绘画	0	24	0	24	书法、绘画
金银锡器	20	0	0	20	金银锡器
乐器、法器	0	0	15	15	乐器、法器
玻璃器	12	0	1	13	玻璃器
首饰	13	0	0	13	首饰
度量衡器	0	0	9	9	度量衡器
外国文物	9	0	0	9	外国文物
其他	0	0	9	9	杂项、其他
成扇	0	8	0	8	成扇
铁器、其他金属	0	0	8	8	铁器、其他金属器
钱币	0	7	0	7	钱币
票据	0	0	6	6	票据
珐琅器	0	0	5	5	珐琅器、珐琅

(To be continued...)

(Continued from previous column)

Original	Gugong	Taipei	Zhejiang	Total	Mapping
古籍图书	0	0	3	3	古籍图书
名人遗物	0	0	2	2	名人遗物
丝绣	0	2	0	2	丝绣
法帖	0	1	0	1	法帖
甲骨	0	0	1	1	甲骨
交通、运输工具	0	0	1	1	交通、运输工具

Table 7: Category mapping statistics across three museums. "Gugong" represents The Palace Museum (Beijing), "Taipei" represents the National Palace Museum, and "Zhejiang" represents the Zhejiang Provincial Museum.

B: Museum-Level Retrieval Performance

Model	T-dim	I-dim	Structural Retrieval					
			Image-to-Text			Text-to-Image		
			R@1	R@5	R@10	R@1	R@5	R@10
CLIP-CN	1024	1024	5.63	15.66	22.03	7.60	18.96	26.00
multimodal-v1	1024	1024	7.25	18.18	25.34	7.70	17.63	24.17
tongyi-flash	768	768	2.65	7.71	11.66	1.98	5.23	7.63
tongyi-plus	1152	1152	4.37	11.32	16.07	3.76	9.20	13.46
qwen2.5-vl	512	512	8.48	20.83	29.00	10.78	22.90	30.10
qwen2.5-vl	768	768	9.13	21.74	29.81	12.24	24.33	31.58
qwen2.5-vl	1024	1024	9.48	22.27	30.52	12.85	25.18	32.35
qwen2.5-vl	2048	2048	**9.94**	**23.18**	**31.21**	**14.14**	**26.83**	**34.82**

Table 8: Performance on the **Structural Retrieval** subtask (Zhejiang subset).

Tables 9 and 10 report the Structural and Descriptive retrieval results on the Zhejiang Provincial Museum subset, respectively. Tables 11 and 12 report the corresponding results on the Palace Museum subset. Tables 13 and 14 report the corresponding results on the National Palace Museum (Taipei) subset.

Model	T	I	Structural Retrieval					
			I→T			T→I		
			R@1	R@5	R@10	R@1	R@5	R@10
CLIP-CN	1024	1024	5.63	15.66	22.03	7.60	18.96	26.00
multimodal-embedding-v1	1024	1024	7.25	18.18	25.34	7.70	17.63	24.17
tongyi-embedding-vision-flash	768	768	2.65	7.71	11.66	1.98	5.23	7.63
tongyi-embedding-vision-plus	1152	1152	4.37	11.32	16.07	3.76	9.20	13.46
qwen2.5-vl-embedding	512	512	8.48	20.83	29.00	10.78	22.90	30.10
qwen2.5-vl-embedding	768	768	9.13	21.74	29.81	12.24	24.33	31.58
qwen2.5-vl-embedding	1024	1024	9.48	22.27	30.52	12.85	25.18	32.35
qwen2.5-vl-embedding	2048	2048	**9.94**	**23.18**	**31.21**	**14.14**	**26.83**	**34.82**

Table 9: Structural retrieval performance (%) on the Zhejiang Provincial Museum subset.

Model	T-dim	I-dim	Descriptive Retrieval					
			I→T			T→I		
			R@1	R@5	R@10	R@1	R@5	R@10
CLIP-CN	1024	1024	4.29	10.37	14.91	6.33	15.08	20.57
multimodal-embedding-v1	1024	1024	6.55	14.57	19.51	11.20	20.02	26.05
tongyi-embedding-vision-flash	768	768	3.58	8.81	12.58	4.64	9.69	13.11
tongyi-embedding-vision-plus	1152	1152	5.50	12.97	17.53	7.48	14.75	19.55
qwen2.5-vl-embedding	512	512	7.31	17.10	23.76	14.73	27.65	34.56
qwen2.5-vl-embedding	768	768	7.42	17.43	24.21	15.48	29.03	35.50
qwen2.5-vl-embedding	1024	1024	7.43	17.73	24.54	16.12	28.97	36.47
qwen2.5-vl-embedding	2048	2048	**7.63**	**18.22**	**24.67**	**17.25**	**30.80**	**38.21**

Table 10: Descriptive retrieval performance (%) on the Zhejiang Provincial Museum subset. Best results are in **bold**.

Model	T-dim	I-dim	Structural Retrieval					
			I→T			T→I		
			R@1	R@5	R@10	R@1	R@5	R@10
CLIP-CN	1024	1024	15.48	36.89	48.02	16.15	37.93	49.48
multimodal-embedding-v1	1024	1024	17.77	37.17	47.70	19.53	38.30	48.84
tongyi-embedding-vision-flash	768	768	7.95	18.67	25.96	4.61	13.41	18.84
tongyi-embedding-vision-plus	1152	1152	11.49	25.98	34.63	8.76	20.96	29.37
qwen2.5-vl-embedding	512	512	17.95	39.03	49.34	15.94	33.13	43.50
qwen2.5-vl-embedding	768	768	18.63	39.74	49.95	18.85	37.55	47.46
qwen2.5-vl-embedding	1024	1024	19.00	40.23	50.52	19.74	38.49	48.97
qwen2.5-vl-embedding	2048	2048	**19.54**	**40.55**	**51.35**	**21.41**	**42.59**	**53.49**

Table 11: Structural retrieval performance (%) on the Palace Museum subset. Best results are in **bold**.

Model	T	I	Descriptive Retrieval					
			I→T			T→I		
			R@1	R@5	R@10	R@1	R@5	R@10
CLIP-CN	1024	1024	15.35	35.60	47.29	15.93	37.74	50.78
multimodal-embedding-v1	1024	1024	19.69	39.18	47.67	24.20	44.50	53.71
tongyi-embedding-vision-flash	768	768	10.20	24.38	32.70	8.70	21.39	29.50
tongyi-embedding-vision-plus	1152	1152	15.70	32.85	41.66	14.94	32.60	42.02
qwen2.5-vl-embedding	512	512	21.46	42.13	51.89	25.11	46.64	57.02
qwen2.5-vl-embedding	768	768	22.08	42.73	53.34	27.73	49.46	59.90
qwen2.5-vl-embedding	1024	1024	22.69	43.80	53.66	28.35	50.93	61.01
qwen2.5-vl-embedding	2048	2048	**22.88**	**43.83**	**53.94**	**30.51**	**52.57**	**62.52**

Table 12: Descriptive retrieval performance (%) on the Palace Museum subset. Best results are in **bold**.

Model	T-dim	I-dim	Structural Retrieval					
			I→T			T→I		
			R@1	R@5	R@10	R@1	R@5	R@10
CLIP-CN	1024	1024	3.88	11.34	17.00	3.68	9.74	14.59
multimodal-embedding-v1	1024	1024	6.65	17.74	24.94	6.56	14.03	19.56
tongyi-embedding-vision-flash	768	768	1.94	6.15	9.46	1.19	3.77	5.93
tongyi-embedding-vision-plus	1152	1152	3.23	9.37	13.70	2.22	6.24	9.36
qwen2.5-vl-embedding	512	512	5.12	14.53	20.91	3.73	10.01	14.81
qwen2.5-vl-embedding	768	768	5.29	14.79	21.23	4.38	11.13	16.36
qwen2.5-vl-embedding	1024	1024	5.51	15.36	21.96	4.39	11.35	16.83
qwen2.5-vl-embedding	2048	2048	**5.71**	**15.85**	**22.51**	**4.89**	**12.65**	**18.50**

Table 13: Structural retrieval performance (%) on the National Palace Museum subset. Best results are in **bold**.

Model	T-dim	I-dim	Descriptive Retrieval					
			I→T			T→I		
			R@1	R@5	R@10	R@1	R@5	R@10
CLIP-CN	1024	1024	3.16	9.59	14.32	4.02	10.16	14.79
multimodal-embedding-v1	1024	1024	5.89	14.73	20.34	7.97	16.23	21.86
tongyi-embedding-vision-flash	768	768	3.02	8.45	12.39	2.33	5.97	8.70
tongyi-embedding-vision-plus	1152	1152	4.74	12.20	17.31	3.91	9.55	13.79
qwen2.5-vl-embedding	512	512	5.76	15.48	21.87	7.49	16.37	22.09
qwen2.5-vl-embedding	768	768	6.19	16.09	22.67	8.20	17.83	23.76
qwen2.5-vl-embedding	1024	1024	6.41	16.62	23.35	8.58	18.29	24.62
qwen2.5-vl-embedding	2048	2048	**6.42**	**16.51**	**23.41**	**8.89**	**19.08**	**25.42**

Table 14: Descriptive retrieval performance (%) on the National Palace Museum subset. Best results are in **bold**.

C: Cultural VQA Prompt Template

The following prompt is used with GPT-4o to generate visual-grounded question–answer pairs during Cultural VQA construction.

Prompt C.1: Visual-Grounded QA Generation

指令：
你是一名视觉问答（VQA）任务的出题助手。输入是一段关于某件文物的文字描述。你的任务是生成尽可能多的高质量问答对（QA pairs）。

规则：

1. 【视觉导向】问题必须围绕可在图像中直接观察或由图像风格合理推断的要素（如形制、结构部件、纹饰、色彩、构图、文字位置、数量、朝向、姿态、整体风格等）。可以涉及年代或作者，但前提是问题明确以视觉特征为依据（例如造型风格、书风、服饰样式等），禁止仅凭文字内容、典故出处或纯粹历史背景即可回答的问题。
2. 【逐点出题】每个问题只针对一个具体视觉点或一个明确的视觉推断维度，不可一次询问多个特征。
3. 【防止数据泄漏】在问题中统一使用“这件文物”“该文物”“此器物”等指代，不得出现文本中已有的具体名称、专有名词或题款原文。
4. 【文本可证】答案必须逐字摘录自输入文本，不得改写、概括或补充推测（包括年代、作者等信息也必须在文本中出现）。
5. 【禁止尺寸相关问题】严格禁止生成任何涉及“尺寸”“高”“宽”“长”“纵”“横”“厚”“大小”“比例”“厘米”“毫米”“开本”“规格”“长度”“高度”等字样的问题。若描述中出现此类信息，应完全忽略。
6. 【多样且充分】生成尽可能多的问题，覆盖不同视觉层面（结构、纹饰、图案、色彩、位置、数量、姿态、风格、可能关联的年代或作者等），但不包括材料产地、流传经历等纯历史信息。
7. 【输出格式】输出为 JSON 数组，每个元素包含：`{"question": "", "answer": ""}`。

示例输入：

“青釉盘口壶，唐，高15厘米，盘口短颈，圆腹，下承圈足，通体施青釉，色泽温润。”

示例输出：

```
[
  {"question": "这件文物的釉色是什么样的？", "answer": "青釉"},
  {"question": "这件文物的腹部形状如何？", "answer": "圆腹"},
  {"question": "这件文物的足部结构是什么？", "answer": "圈足"},
  {"question": "这件文物的颈部特征是什么？", "answer": "盘口短颈"}
]
```

Prompt C.1: Visual-Grounded QA Generation (English)

Instruction:
You are an assistant for creating questions for Visual Question Answering (VQA). The input is a textual description of a museum artifact. Your task is to generate as many high-quality QA pairs as possible.

Rules:

1. **[Visually grounded]** Each question must focus on factors that can be directly observed in the image or reasonably inferred from visual style (e.g., form, structural components, patterns, colors, composition, location of inscriptions, counts, orientation, pose, overall style). Questions may involve period or author only if they are explicitly grounded in visual evidence (e.g., stylistic features, calligraphic style, clothing style). Questions that can be answered purely from textual content, allusions, or historical background are forbidden.
2. **[One point per question]** Each question should target exactly one specific visual point or one clear dimension of visual inference. Do not ask about multiple attributes in a single question.
3. **[Prevent leakage]** In questions, refer to the artifact only as "this artifact", "the artifact", "this object", etc. Do not include any specific names, proper nouns, or the original inscription text that appears in the input description.
4. **[Text-verifiable]** Each answer must be copied verbatim from the input text. Do not paraphrase, summarize, or add any speculation (including period/author information, which must also appear in the input text).
5. **[No size-related questions]** It is strictly forbidden to generate any questions involving size, such as "dimensions", "height", "width", "length", "thickness", "scale", "cm", "mm", "format", or "specification". If such information appears in the description, ignore it completely.
6. **[Diverse and sufficient]** Generate as many questions as possible, covering different visual aspects (structure, ornamentation, motifs, colors, position, quantity, pose, style, and visually grounded cues that may relate to period or author), but exclude purely historical information such as provenance or circulation history.
7. **[Output format]** Output a JSON array. Each element should be: `{"question": "", "answer": ""}`.

Example input:

"A celadon-glazed vase with a dish-shaped mouth, Tang dynasty, height 15 cm. It has a dish-shaped mouth and short neck, a rounded belly, and a ring foot. The whole body is covered with celadon glaze, with a smooth and lustrous tone."

Example output:

```
[
  {"question": "What kind of glaze color does this artifact have?", "answer": "celadon glaze"},
  {"question": "What is the shape of this artifact's belly?", "answer": "rounded belly"},
  {"question": "What is the structure of the foot of this artifact?", "answer": "ring foot"},
  {"question": "What are the features of this artifact's neck?", "answer": "dish-shaped mouth and short neck"}
]
```

Prompt C.2: Visual-Grounded QA Verification

指令：
你是一名博物馆视觉问答（VQA）数据的质量核验员。输入包括一段文物的文字描述，以及若干已生成的问题–答案对（QA pairs）。其中，答案被视为来自专家文本的参考答案。你的任务是对每一个问答对进行逐条核验，判断该问题是否清晰可理解，并且能够被给定答案与输入文本合理支撑。

核验标准：

1. 【视觉可理解性】问题应围绕可从图像中直接观察，或可依据视觉风格合理推断的要素（如形制、结构、纹饰、色彩、构图、数量、位置或整体风格）。
2. 【答案一致性】给定答案必须在输入文本中有明确、逐字可对应的依据，且问题应与该答案形成清晰的一一对应关系。
3. 【单一焦点】每个问题只针对一个明确的视觉或语义要点，不得混合询问多个属性。
4. 【信息隔离】问题中不得出现文物的具体名称、专有名词或直接复现文本中的原句或题款内容。
5. 【排除非视觉问题】涉及尺寸、精确数值、或纯文本类历史背景的信息，应判定为不合格。
6. 【表达清晰】问题表述应清楚、无歧义，不影响对给定答案的理解；允许使用专业术语，但不得引入额外信息。

输出要求：

- 对每一个问答对输出一个核验结果。
- 使用字段 `valid` 标注该问答是否合格，取值为 `true` 或 `false`。
- 若 `valid = false`，请从下列原因中选择一个最主要的原因填入 `reason` 字段：
  - `not_visual`（问题无法从视觉信息理解）
  - `multi_aspect`（问题涉及多个要点）
  - `information_leakage`（问题泄漏文物名称或原文内容）
  - `unclear_question`（问题表述不清晰或存在歧义）
- 若 `valid = true`，`reason` 设为 `null`。
- 不要输出任何额外解释性文字。

输出格式：

```
[
  {
    "question": "",
    "answer": "",
    "valid": true,
    "reason": null
  },
  {
    "question": "",
    "answer": "",
    "valid": false,
    "reason": "multi_aspect"
  }
]
```

Prompt C.2: Visual-Grounded QA Verification (English)

Instruction:
You are a quality auditor for museum Visual Question Answering (VQA) data. The input includes a textual description of an artifact and several generated question-answer (QA) pairs. Here, the **answers are treated as reference answers from expert text**. Your task is to verify **each QA pair** one by one, and determine whether the **question is clear and understandable, and whether it can be reasonably supported by the given answer and the input text**.

Verification criteria:

1. **[Visual interpretability]** The question should focus on factors that can be directly observed in the image or reasonably inferred from visual style (e.g., form, structure, ornamentation/patterns, colors, composition, quantity, position, or overall style).
2. **[Answer consistency]** The given answer must have explicit, word-by-word support in the input text, and the question should form a clear one-to-one correspondence with that answer.
3. **[Single focus]** Each question should target only one clear visual or semantic point, and must not mix multiple attributes.
4. **[Information isolation]** The question must not include the artifact's specific name, proper nouns, or directly reproduce sentences from the input text or any inscription content.
5. **[Exclude non-visual questions]** Questions involving dimensions, exact numeric values, or purely textual/historical background information should be marked as invalid.
6. **[Clarity]** The question should be clearly phrased and unambiguous, and should not hinder understanding of the given answer; professional terminology is allowed, but no additional information may be introduced.

Output requirements:

- Output one verification result for **each QA pair**.
- Use the field `valid` to indicate whether the QA pair is valid, with value `true` or `false`.
- If `valid = false`, choose **one primary reason** from the following and fill it into the `reason` field:
  - `not_visual` (the question is not interpretable from visual information)
  - `multi_aspect` (the question covers multiple aspects)
  - `information_leakage` (the question leaks the artifact name or original text content)
  - `unclear_question` (the question is unclear or ambiguous)
- If `valid = true`, set `reason` to `null`.
- Do not output any additional explanatory text.

Output format:

```
[
  {
    "question": "",
    "answer": "",
    "valid": true,
    "reason": null
  },
  {
    "question": "",
    "answer": "",
    "valid": false,
    "reason": "multi_aspect"
  }
]
```