MUSEFACET: BENCHMARKING TRUSTWORTHY AND CONTROLLABLE IMAGE-TO-TEXT GENERATION FOR MUSEUM COLLECTION INTERPRETATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Museum collection interpretation represents a challenging task for Multimodal Large Language Models (MLLMs), requiring accurate object identification and knowledge-grounded, context-aware generation beyond literal visual descriptions. However, existing efforts on open-ended text generation for artworks and museum objects remain limited in dataset scale and diversity or in evaluation protocols that rely largely on reference-based similarity metrics. To address these limitations, we introduce MuseFACET, a benchmark comprising 5,000 museum objects from 9 museums. MuseFACET presents two hierarchical tasks: Identification task for predicting object attributes from images and Interpretation task for generating coherent, context-aware descriptions. We further propose an evaluation framework emphasizing trustworthiness and controllability through identification accuracy and five interpretation criteria: Faithfulness, Completeness, Conciseness, Depth, and Acuity. Evaluation of 14 proprietary and open-source MLLMs shows that the highest accuracy on the Identification task reaches only 31.2%. For Interpretation, models exhibit persistent limitations in factual grounding and object-specific insight under the image-only setting, while providing metadata substantially improves performance. Model performance also varies across cultural regions, revealing uneven capabilities in cultural interpretation. MuseFACET provides a systematic benchmark for museum collection interpretation and a concrete evaluation protocol for trustworthy and controllable open-ended image-to-text generation.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023a) have made rapid progress in visual understanding and vision-to-language generation. However, trustworthy and controllable vision-to-language generation remains challenging, as models may generate unsupported content or fail to satisfy expected generation requirements (Li et al., 2023b; Jing et al., 2024; Liang et al., 2024). Museum collection interpretation provides a representative setting for addressing these challenges, playing an important role in communicating cultural heritage by making museum objects and their historical and cultural significance accessible to diverse audiences. It requires models to accurately identify visual details, such as historical or mythological figures, and generate coherent interpretive descriptions grounded in reliable object information (Balauca et al., 2025; Hayashi et al., 2024). Beyond conventional image captioning, this task demands both fine-grained visual understanding and context-aware generation that conveys an object's background, meaning, and significance in accordance with professional interpretive standards.

Existing vision-language benchmarks predominantly evaluate models through constrained response formats, such as multiple-choice questions and short-answer visual question answering, while open-ended generation remains less extensively studied. This tendency can also be observed in culture-focused or museum-grounded benchmarks, where such constrained evaluation frameworks are commonly used to assess models' cultural understanding from visual inputs (Garcia et al., 2020; Nayak et al., 2024; Balauca et al., 2025; Zhang et al., 2026). While these frameworks provide useful measures of cultural knowledge and reasoning, real-world applications also require the ability to generate open-ended text. Accordingly, a complementary line of research has begun to study open-ended

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

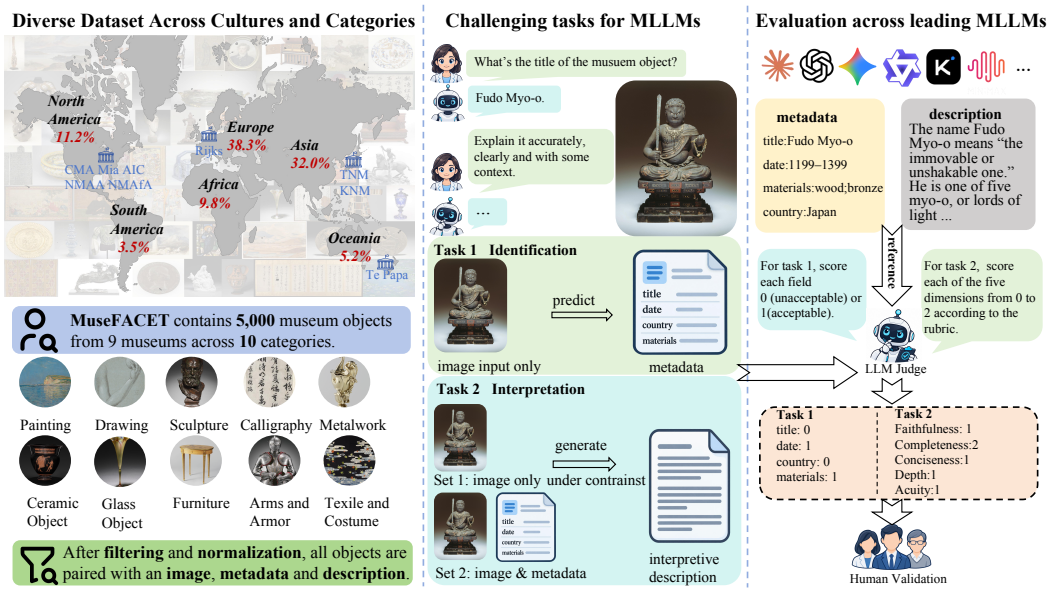


Figure 1: Overview of MuseFACET. Left: 5,000 museum collections from 9 museums across 10 categories and diverse geographic regions. Middle: Identification predicts collection metadata, while Interpretation generates descriptions under image-only and metadata-guided settings. Right: models are evaluated with field-level scoring for Identification and five-dimensional rubric-based scoring for Interpretation, complemented by human validation.

generation, including knowledge-informed artwork descriptions, structured expert-style interpretation and reinterpretation, and the interpretation of historical and cultural artifacts (Bai et al., 2021; Hayashi et al., 2024; Ghaboura et al., 2025; Wei et al., 2026).

However, existing studies still face two key limitations. First, there is a lack of large-scale datasets with diversity in both object categories and cultural regions for museum collection interpretation, making it difficult to evaluate models across heterogeneous museum collections(Bai et al., 2021; Hayashi et al., 2024; Ghaboura et al., 2025; Li et al., 2026; Wei et al., 2026). Second, existing evaluation protocols still rely heavily on traditional reference-based metrics(Bai et al., 2021; Hayashi et al., 2024; Ghaboura et al., 2025; Wei et al., 2026), such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), to measure similarity to human-written descriptions. Such metrics may fall short for artwork or museum object interpretation, as they provide limited insight into specific errors or fine-grained differences in generation quality and may correlate poorly with human judgments(Bai et al., 2021; Hessel et al., 2021; Anderson et al., 2016).

Accordingly, we introduce MuseFACET, a benchmark for comprehensively evaluating museum collection interpretation in MLLMs. MuseFACET comprises 5,000 objects from 9 museums, spanning 10 object categories and diverse geographic origins, with each object paired with an image, normalized metadata, and a museum-provided description. Our evaluation emphasizes two key aspects: **Trustworthiness**, which calls for generated interpretations to be grounded in reliable information rather than unsupported inferences, and **Controllability**, which requires interpretations to meet specified content and presentation requirements. To assess both aspects, MuseFACET includes two hierarchical tasks. **Identification** task evaluates the recognition of fundamental object attributes, including title, date, materials, and country, from images, while **Interpretation** task assesses the generation of coherent descriptions in image-only and metadata-guided settings. Following task requirements and professional museum practices, interpretations are evaluated along five dimensions: Faithfulness, Completeness, Conciseness, Depth, and Acuity. Figure 1 summarizes the dataset, task design, and evaluation pipeline.

We evaluate 14 proprietary and open-source MLLMs on MUSEFACET. For Identification, the strongest model correctly predicts all four attributes for only 31.2% of objects. For Interpretation, models struggle to generate grounded, specific, and concise descriptions from images alone, while metadata improves the average score of proprietary models from 4.58 to 7.00, highlighting its im-

Table 1: Comparison with representative benchmarks for open-ended artwork and cultural interpretation. Desc. Src.: description source; Geo. Cov.: geographic coverage; Meta.: metadata-guided input; Multi-dim. Eval.: multi-dimensional evaluation. ✓, △, and ×denote full, partial/restricted, and absent coverage, respectively. Geographic coverage follows the units reported by each work; TimeTravel reports historical regions. "LLM + Expert" denotes LLM-generated descriptions subsequently refined by domain experts.

	Dataset		Task		Setting		Eval.
Benchmark	**Desc. Src.**	**Geo. Cov.**	**Ident.**	**Interp.**	**Meta. Interp.**	**Web**	**Multi-dim. Eval.**
ArtEmis (Achlioptas et al., 2021)	Crowd	–	×	△	×	×	×
Explain Me the Painting (Bai et al., 2021)	SemArt	Europe	×	✓	×	×	✓
Towards Artwork Explanation (Hayashi et al., 2024)	Wikipedia	–	×	✓	△	×	×
TimeTravel (Ghaboura et al., 2025)	LLM + Expert	10 regions	△	✓	×	×	△
RAVENEA (Li et al., 2026)	CCUB	8 countries	×	△	×	×	×
CArtBench (Wei et al., 2026)	Expert	China	✓	✓	×	×	✓
MuseFACET (Ours)	**Museum**	**6 cont./89 countries**	✓	✓	✓	✓	✓

portance for reliable generation. Cultural analysis further reveals regional performance disparities, with relatively stronger results in Europe and North America than in Africa, South America, and Oceania.

Our contributions are threefold[1]. First, we construct a diverse museum benchmark comprising 5,000 objects from 9 museums across a wide range of object categories and geographic origins, with object images, manually normalized and verified metadata, and museum-provided descriptions. Second, we introduce two hierarchical tasks, Identification and Interpretation, with five complementary criteria to assess the trustworthiness and controllability of open-ended generation. Third, we evaluate 14 MLLMs, analyzing the relationship between Identification and Interpretation, the effects of metadata and web search, and performance variation across cultural regions.

2 Related Works

Vision-Language Benchmarks. Vision-language benchmarks have progressed from image captioning and visual question answering (Chen et al., 2015; Antol et al., 2015) toward tasks that require broader contextual knowledge and reasoning. Subsequent benchmarks extend visual understanding beyond directly observable content by incorporating commonsense or external knowledge (Zellers et al., 2019; Marino et al., 2019; Schwenk et al., 2022). More recent benchmarks further broaden the scope of evaluation, covering multiple aspects of perception, knowledge, and reasoning within unified evaluation suites (Fu et al., 2025; Liu et al., 2024b; Yu et al., 2023).

Despite this progress, many widely used vision-language benchmarks still rely on multiple-choice questions, short-form VQA, or other constrained response formats (Antol et al., 2015; Fu et al., 2025; Liu et al., 2024b). Such evaluations do not directly assess model behavior in open-ended generation, where fluent responses may still contain hallucinated or visually unsupported content (Li et al., 2023b; Jing et al., 2024). Beyond factual reliability, open-ended generation further raises the question of controllability. Controllable text generation requires outputs to satisfy predefined control conditions while maintaining generation quality (Liang et al., 2024). Accordingly, evaluating open-ended generation requires considering not only whether the generated content is faithful to the input, but also whether it conforms to the desired content and generation constraints.

Resources and Benchmarks for Museum and Cultural Heritage. Museums and digital archives provide valuable resources for studying visual culture, as their digitized collections often combine object images with structured catalog metadata, curatorial descriptions, and historical context. Museums have also developed established practices for documenting and describing cultural objects. Professional guidelines such as the Categories for the Description of Works of Art (CDWA) and Cataloging Cultural Objects (CCO) provide guidance on the organization, cataloging, and description of collection information (Baca & Harpring, 2017; Baca, 2006), offering useful domain references for museum-oriented dataset construction and evaluation.

[1] All the code and dataset will be released upon acceptance.

Recent work has begun to benchmark multimodal models directly in the domains of art, museum collections, and cultural heritage. Many existing evaluations, however, are formulated primarily as visual question answering, short-answer prediction, or other structured tasks (Garcia et al., 2020; Nayak et al., 2024; Balauca et al., 2025; Zhang et al., 2026; Mensink & Van Gemert, 2014). A smaller body of work considers open-ended description or interpretation, but the corresponding datasets are often limited in scale or collection scope (Bai et al., 2021; Hayashi et al., 2024; Ghaboura et al., 2025; Li et al., 2026; Wei et al., 2026). These studies commonly evaluate generated text using reference-based metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), and BERTScore (Zhang et al., 2019), alongside other automatic or model-based evaluation methods (Bai et al., 2021; Hayashi et al., 2024; Ghaboura et al., 2025; Wei et al., 2026).Such metrics primarily measure similarity to reference texts and provide limited insight into whether a generated interpretation is factually grounded, sufficiently informative, concise, and specific. Table 1 summarizes the main differences between MuseFACET and representative benchmarks for artwork and cultural interpretation.

3 MUSEFACET

3.1 DATA COLLECTION AND PREPROCESSING

Figure 2 summarizes the overall data construction, annotation, and quality-control pipeline.

Data collection. To ensure dataset diversity, we emphasize broad coverage of museum-object categories and geographic origins, as illustrated in Figure 1. We collect objects from 9 museums, with objects originating from 89 countries across six continents: *Africa, Asia, Europe, North America, South America, and Oceania*. Our selection is guided by three considerations. First, collecting objects from museums in different regions helps enhance cultural diversity and better represent museum collections worldwide. Second, these museums offer publicly accessible digital collections with clear terms of use, enabling data acquisition through official APIs. Third, they provide the images, metadata, and descriptions required for our Identification and Interpretation tasks. Detailed information about the data sources and collection procedures is provided in Appendix A.

Schema alignment. Since museum cataloging conventions vary across institutions, records from different museums need to be aligned into a common structured framework (Doerr, 2003). We define a shared schema comprising *category*, *title*, *date*, *country*, *materials*, *artist*, *description*, and image information. These fields reflect information commonly presented to the public in museum galleries and online collection pages, and are consistent with established professional guidance such as Cataloging Cultural Objects (CCO) and the Categories for the Description of Works of Art (CDWA) (Baca, 2006; Baca & Harpring, 2017). We then map museum-specific fields to this shared schema. Further details on schema alignment are provided in Appendix B.

Rule-based pre-annotation. Following schema alignment, we apply field-specific and source-specific rules to normalize metadata values across museums in terms of granularity, terminology, and format. The normalization focuses on category, country, date, and materials. For category and country, museum-specific labels are mapped through manually curated mapping tables to the ten shared object categories used in MuseFACET and standardized country-level labels, respectively. For date and materials, we apply rule-based parsing and extraction to convert heterogeneous source values into a year or year range and a standardized list of principal materials, respectively. Detailed normalization rules, mappings, and additional examples are provided in Appendix C.

Diversity-oriented sampling. To construct a diverse and tractable benchmark, we perform stratified sampling over the joint distribution of category and country, treating each category–country combination as a stratum. Category and country are pre-annotated using manually curated mapping tables, yielding sufficiently reliable labels for stratified sampling, as further supported by their low subsequent correction rates. To mitigate the distributional imbalance in the source collections and increase the representation of underrepresented combinations, we adopt square-root allocation, which assigns relatively greater sampling weight to less frequent strata while reducing the dominance of frequent ones.We further use a greedy selection procedure that prioritizes previously uncovered category–country combinations, thereby increasing joint coverage under a fixed sampling budget. Detailed sampling procedures are provided in Appendix D.

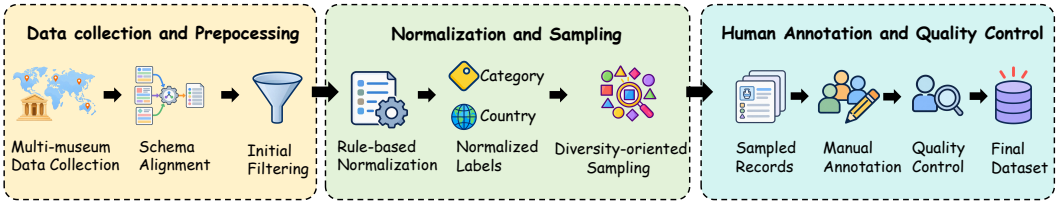


Figure 2: Overview of the MuseFACET construction pipeline.

3.2 Human Annotation and Quality Control

Metadata annotation. Since rule-based pre-annotation cannot accurately handle all cases and can still produce errors, we recruit 9 annotators to further review and correct the sampled records. After receiving training on the annotation guidelines, annotators complete a pilot annotation of 300 records. We then discuss ambiguous cases and refine the guidelines before formal annotation. Each record is independently reviewed by one annotator through a dedicated web-based interface displaying the object image, raw metadata, pre-annotations, and museum-provided description. Annotators assess record-level consistency, metadata fidelity, and description quality using eight predefined binary check items. Identified errors are flagged and corrected based on original museum evidence, while ambiguous cases are reserved for subsequent quality control. Overall, 18.8% of the sampled records are flagged with at least one issue for subsequent correction. Detailed annotation criteria, correction statistics, and the verification interface are provided in Appendix E.

Quality control. Three reviewers who developed the annotation guidelines conduct the quality-control process. They jointly review all flagged records and the corrections proposed by annotators, with particular attention to cases where the annotator is uncertain about the appropriate correction. The reviewers discuss these cases and determine the final values based on the original museum evidence. Records that cannot be resolved from the available museum evidence are removed from the dataset. At the final stage, we manually review, filter, and correct the dataset again, resolving remaining metadata inconsistencies and correcting formatting or language errors. Since quality control may remove some records, additional samples are drawn from the candidate pool to maintain the target dataset size and distribution, and undergo the same annotation and quality-control procedure.

3.3 Dataset Statistics

Broad spatiotemporal coverage.

As shown in Figure 3, MuseFACET draws from nine museums and covers six continents and 89 country-level origins. The dataset also spans a broad historical range from approximately 4000 BCE to 2022 CE. This broad geographic and temporal coverage supports evaluation across diverse cultural contexts and historical periods.

Diverse collection categories. As shown in Figure 3, MuseFACET covers 10 collection categories together with a wide range of fine-grained types. These span traditional fine-art categories such as painting, drawing, and sculpture, as well as decorative arts including ceramics, furniture, textiles, glass, metalwork, and arms and armor.

High-quality metadata and descriptions. Each object in MuseFACET includes manually verified task-relevant metadata, an accessible image, and a museum-provided description. As summarized in Table 2, titles contain 4.79 words and materials contain 2.00 entries on average, providing concise targets for Identification. The descriptions average 121.2 words, with a median of 99 words, providing richer contextual information for Interpretation.

Statistic	Value
Total objects	5,000
Country-level origins	89
Collection categories	10
Title length (words)	
Mean / Median	4.79 / 4
Materials per collection	
Mean / Median	2.00 / 2
Date (interval midpoint)	
$\leq$1500 CE	993 (19.9%)
1501–1800	1,855 (37.1%)
>1800	2,152 (43.0%)
Description length (words)	
Mean / Median	121.2 / 99
20–99	2,513 (50.3%)
100–199	1,843 (36.9%)
$\geq$200	644 (12.9%)

Table 2: Statistics of MuseFACET.

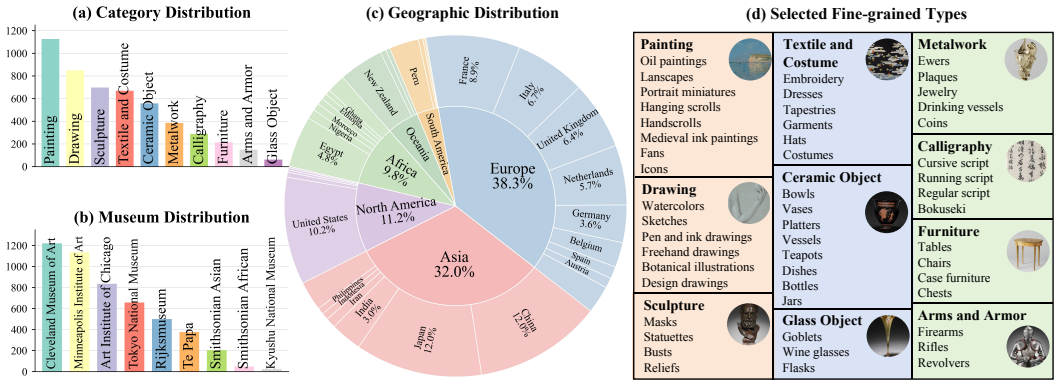


Figure 3: MuseFACET composition. (a) Object-category distribution; (b) source-museum distribution; (c) geographic distribution by continent and country; and (d) representative fine-grained object types within the ten categories.

MuseFACET-testmini. Alongside the full 5,000-object benchmark, we provide *MuseFACET-testmini*, a 1,000-object subset for efficient evaluation. We construct the subset to closely preserve the joint category–country distribution of the full dataset, retaining its main compositional characteristics while substantially reducing evaluation cost. *MuseFACET-testmini* therefore provides a lightweight option for rapid model development and resource-constrained evaluation. More detailed dataset statistics, cross-source distributions, and representative examples are provided in Appendix F.

3.4 Task Design

We design two hierarchical tasks, Identification and Interpretation, to evaluate factual inference and open-ended text generation separately. The complete generation prompts and output formats are provided in Appendix G.

For **Identification**, models are given an image of a museum object and asked to predict four attributes in JSON format: title, date, materials, and country. These attributes are available across all selected museums and capture fundamental aspects of museum cataloging, including object identity, physical characteristics, chronology, and geographic origin (Baca & Harpring, 2017; Baca, 2006).

For **Interpretation**, models generate an open-ended description of a museum object under two input settings. The image-only setting provides only the object image, while the metadata-guided setting additionally provides its title, date, materials, country, and artist information. In both settings, models generate a concise, coherent paragraph describing relevant visual, artistic, historical, and cultural characteristics. By comparing the two settings, we examine how access to reliable object information affects subsequent interpretation.

4 Experiments

4.1 Experimental Setup

Models: We evaluate 14 MLLMs, including both proprietary and open-source models. For proprietary models, we test Claude Opus 5, Gemini 3.1 Pro Preview, Kimi K2.6, Qwen 3.7 Plus, GLM-5V Turbo (Team et al., 2026b), Doubao Seed 2.0 Pro (Seed, 2026), and MiniMax M3. For open-source models, we report results for Gemma-4-31B-IT (Team et al., 2026a), GLM-4.1V-9B-Thinking (Hong et al., 2025), NVLM-D-72B (Dai et al., 2024), Qwen3-VL-30B-A3B (Bai et al., 2025), MiniCPM-V-4.5 (Yu et al., 2026), Llama-4-Scout, and InternVL3.5-8B (Wang et al., 2025). Detailed model information and inference settings are provided in Appendix H.

Settings: Identification is evaluated in an image-only setting, where models predict normalized metadata directly from the input image. Interpretation is evaluated under two settings: image-only and metadata-guided, where the latter additionally provides the title, date, materials, country, and

Table 3: Identification results on MuseFACET and MuseFACET-testmini. All is the percentage of samples with all four fields correct. Bold and underlined values denote the best and second-best results within each model group.

Model	MuseFACET					MuseFACET-testmini				
	All	Title	Date	Mat.	Country	All	Title	Date	Mat.	Country
Proprietary MLLMs										
Claude Opus 5	**31.22**	**64.44**	**67.88**	86.14	**70.44**	**30.31**	**64.41**	**63.89**	86.06	**70.54**
Kimi K2.6	24.12	60.26	62.60	85.99	65.93	24.00	59.09	60.36	86.47	67.23
Doubao Seed 2.0 Pro	22.74	60.44	58.52	**88.78**	64.70	22.60	61.99	53.75	**89.33**	65.05
Gemini 3.1 Pro Preview	17.98	54.04	51.46	84.26	67.12	14.99	50.79	46.36	82.37	67.48
GLM-5V Turbo	15.04	50.44	54.82	82.68	58.60	15.73	53.03	52.50	84.80	57.81
Qwen 3.7 Plus	14.95	47.21	54.31	82.42	55.98	14.59	46.72	50.11	83.83	56.13
MiniMax M3	12.81	43.02	53.07	83.54	55.57	13.54	47.02	49.40	83.64	56.88
Open-Source MLLMs										
Qwen3-VL-30B-A3B	**6.60**	33.40	**47.20**	58.00	**52.60**	**8.70**	36.40	**61.90**	63.00	**56.00**
Gemma-4-31B-IT	5.90	**35.00**	45.60	**60.40**	49.90	7.90	**36.60**	59.00	**66.10**	53.10
MiniCPM-V-4.5	3.90	28.00	47.10	50.30	47.60	6.00	33.50	56.20	54.80	53.60
GLM-4.1V-9B-Thinking	3.90	32.00	39.20	52.70	47.60	5.60	27.70	52.40	56.00	52.30
NVLM-D-72B	2.40	22.90	33.50	48.00	38.40	2.40	21.40	44.40	47.40	36.10
Llama-4-Scout	2.10	23.30	28.40	54.10	44.70	5.20	23.40	55.30	55.40	51.30
InternVL3.5-8B	0.20	14.10	8.60	38.50	22.00	1.60	23.20	23.20	46.10	30.70

artist information described in Section 3.2. We also evaluate a subset of proprietary models using their official web-search functionality and report the results separately.

Evaluation: For Identification, we evaluate title, country, and materials using semantic matching rather than exact string matching, with each field scored independently as either 0 (*unacceptable*) or 1 (*acceptable*) (Liu et al., 2023b; Zheng et al., 2023). For Interpretation, we develop an evaluation protocol informed by established guidelines and recommendations for museum and cultural-heritage interpretation. Based on recurring principles identified across these sources, we evaluate each generated description along five complementary dimensions: **Faithfulness** concerns whether its claims are consistent with the depicted collection, the provided metadata, and museum-provided information. **Completeness** captures whether the core information needed to understand the specific collection is sufficiently covered. **Conciseness** concerns clarity and focus, penalizing unnecessary repetition, tangential content, and excessive detail. **Depth** reflects the extent to which the description goes beyond identification and surface-level observation to provide substantive artistic, historical, cultural, or functional analysis. **Acuity** captures the specificity and appropriateness of the interpretation, distinguishing collection-specific insights from generic observations or weakly supported interpretations. Each dimension is scored independently as 0 (*poor*), 1 (*acceptable*), or 2 (*excellent*), yielding a total score from 0 to 10. Detailed scoring criteria, judge prompts, and judge configuration are provided in Appendix I. We further examine evaluator sensitivity and judge self-evaluation bias in Appendix J, and provide representative qualitative error cases in Appendix K.

4.2 Main Results

All models struggle with exact Identification. As shown in Table 3, on MuseFACET, the best-performing model, Claude Opus 5, correctly predicts all four fields for only 31.22% of the objects. The strongest open-source model, Qwen3-VL-30B-A3B, achieves only 6.60%, which is still 6.21 percentage points below MiniMax M3, the weakest proprietary model. This indicates a substantial gap between proprietary and open-source models in visual recognition and reasoning for complex museum objects. A field-level breakdown further shows that models perform considerably better on *materials* than on the other three fields, with all proprietary models achieving over 82% accuracy. In contrast, most models perform worst on *title*. These results suggest that models perform better on relatively constrained visual identification, while open-ended identification requiring knowledge and contextual information remains more challenging.

Generated interpretation remains difficult to make trustworthy and controllable. Under the image-only setting, all models obtain relatively low overall scores, particularly on *Faithfulness*,

Table 4: Interpretation results under image-only and metadata-guided settings. Total is the sum of five 0–2 dimension scores. Bold and underlined values denote the best and second-best results within each model group.

Model	Image Only						Metadata Guided					
	Total	**Faith.**	**Comp.**	**Conc.**	**Depth**	**Acuity**	**Total**	**Faith.**	**Comp.**	**Conc.**	**Depth**	**Acuity**
Proprietary MLLMs												
Claude Opus 5	5.211	1.052	**1.180**	0.466	**1.731**	0.781	**7.700**	1.584	**1.754**	0.909	**1.983**	**1.469**
Gemini 3.1 Pro Preview	**5.634**	**1.164**	1.146	**1.091**	1.424	**0.809**	7.490	**1.672**	1.555	1.285	1.688	1.290
Qwen 3.7 Plus	4.254	0.752	0.861	0.931	1.281	0.429	6.868	1.451	1.362	**1.296**	1.671	1.088
GLM-5V Turbo	4.256	0.716	0.870	0.892	1.346	0.431	6.841	1.422	1.372	1.241	1.714	1.092
Doubao Seed 2.0 Pro	4.184	0.731	1.028	0.466	1.446	0.514	6.854	1.424	1.556	0.849	1.839	1.187
Kimi K2.6	4.556	0.802	0.997	0.564	1.615	0.578	6.712	1.399	1.417	0.885	1.913	1.098
MiniMax M3	3.971	0.806	0.891	0.677	1.197	0.400	6.562	1.349	1.436	1.007	1.679	1.090
Open-Source MLLMs												
Gemma-4-31B-IT	**4.263**	0.810	**0.826**	**1.182**	**1.014**	**0.431**	**6.463**	1.581	1.033	**1.596**	**1.246**	**1.007**
Qwen3-VL-30B-A3B	3.600	0.648	0.748	1.004	0.904	0.296	5.840	1.429	**1.035**	1.208	1.197	0.971
MiniCPM-V-4.5	3.536	0.668	0.676	0.972	0.946	0.274	5.845	1.522	1.006	1.331	1.050	0.936
GLM-4.1V-9B-Thinking	3.196	**0.839**	0.638	0.570	0.847	0.302	5.214	1.701	1.001	0.519	1.056	0.937
NVLM-D-72B	2.663	0.462	0.364	1.126	0.568	0.143	5.914	1.657	0.997	1.296	1.023	0.941
Llama-4-Scout	2.674	0.474	0.438	0.865	0.747	0.150	5.810	**1.732**	1.007	1.053	1.064	0.954
InternVL3.5-8B	2.849	0.630	0.492	0.731	0.826	0.170	5.437	1.490	0.976	1.068	1.017	0.886

Table 5: Web-search results for four proprietary models. Img. and Meta. denote image-only and metadata-guided settings, respectively. Parentheses show changes from paired no-search baselines; bold and underlined values mark the best and second-best results. Full no-search results are in Appendix L.

Model	Identification					Set.	Interpretation					
	All	**Title**	**Date**	**Mat.**	**Country**		**Total**	**Faith.**	**Comp.**	**Conc.**	**Depth**	**Acuity**
Kimi K2.6	**25.00** (+1.00)	59.40 (+0.31)	**61.70** (+1.34)	84.00 (-2.47)	**66.00** (-1.23)	*Img.*	4.042 (-0.247)	0.579 (-0.089)	1.094 (+0.129)	0.191 (-0.384)	**1.643** (+0.102)	0.535 (-0.004)
						Meta.	6.477 (+0.124)	1.089 (-0.141)	1.782 (+0.443)	0.467 (-0.439)	**1.972** (+0.135)	1.167 (+0.127)
Qwen 3.7 Plus	14.21 (-0.37)	49.45 (+2.73)	48.75 (-1.36)	80.68 (-3.15)	57.16 (+1.03)	*Img.*	4.326 (+0.440)	0.722 (+0.124)	0.875 (+0.081)	**0.881** (-0.038)	1.362 (+0.181)	0.486 (+0.092)
						Meta.	**6.743** (+0.324)	**1.412** (+0.131)	1.353 (+0.122)	**1.188** (-0.183)	1.726 (+0.215)	1.063 (+0.039)
Doubao Seed 2.0 Pro	24.50 (+1.90)	**64.70** (+2.71)	53.10 (-0.65)	**86.20** (-3.13)	65.70 (+0.65)	*Img.*	**4.624** (+0.613)	**0.747** (+0.148)	**1.136** (+0.098)	0.577 (+0.089)	1.542 (+0.158)	**0.622** (+0.120)
						Meta.	6.493 (+0.216)	1.186 (-0.018)	1.596 (+0.191)	0.698 (-0.167)	1.902 (+0.166)	1.111 (+0.044)
MiniMax M3	17.43 (+3.89)	46.29 (-0.73)	50.00 (+0.60)	81.26 (-2.38)	58.12 (+1.24)	*Img.*	3.523 (-0.192)	0.525 (-0.178)	0.960 (+0.075)	0.307 (-0.354)	1.370 (+0.269)	0.362 (-0.005)
						Meta.	6.510 (+0.384)	1.076 (-0.101)	**1.836** (+0.470)	0.467 (-0.547)	1.924 (+0.365)	**1.207** (+0.198)

Conciseness, and *Acuity* (shown in Table 4). This suggests that visual information alone makes it difficult to generate interpretations that are well grounded and sufficiently specific while remaining concise and clear. When metadata is provided, all models improve substantially (Table 4; Figure 4), indicating that reliable object information provides an important grounding signal for Interpretation.

4.3 Human Evaluation

Human validation. To validate the reliability of our evaluation protocol, we conduct human evaluation on 200 model responses, stratified across object categories, cultural regions, and model configurations. The sample covers all seven proprietary models and both Interpretation settings. Each response is independently rated by three annotators using the same five-

Table 6: Human–human and human–LLM agreement across the five Interpretation dimensions and the 0–10 Total score. H–H uses Krippendorff's α and H–LLM uses Spearman's ρ.

Dimension	H–H α	H–LLM ρ
Total	**0.833**	**0.758**
Faith.	0.761	0.679
Completeness	0.685	0.622
Conciseness	0.666	0.546
Depth	0.648	0.531
Acuity	0.662	0.641

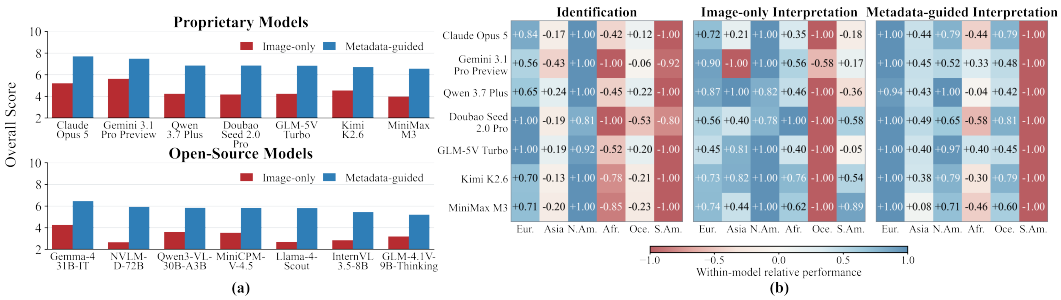


Figure 4: Interpretation performance and cross-cultural variation. (a) Overall Interpretation scores of proprietary and open-source MLLMs under image-only and metadata-guided settings. (b) Relative performance across cultural regions for Identification, image-only Interpretation, and metadata-guided Interpretation, after within-model and within-task min–max normalization to $[-1, 1]$.

dimensional 0–2 rubric as the automatic evaluator. Model identities are hidden and responses are presented in randomized order. Full sampling details, annotator instructions, qualifications, and interface design are provided in Appendix M. We report ordinal Krippendorff's α (Hayes & Krippendorff, 2007) for inter-annotator agreement and Spearman's ρ between automatic scores and mean human ratings. As shown in Table 6, α ranges from 0.648 to 0.761 across dimensions and reaches 0.833 for the Total score, while Human–LLM ρ ranges from 0.531 to 0.679 and reaches 0.758 overall. These results support the reliability of the rubric and its automatic evaluation.

4.4 Findings and Analysis

Model performance varies systematically across cultural regions. We compare relative model performance across six continent-level cultural regions: Europe, Asia, North America, Africa, South America, and Oceania. To make regional patterns comparable across tasks with different score ranges, we compute the mean performance of each model within each region and normalize these regional scores separately for each model and task. The resulting values reflect relative strengths and weaknesses across regions rather than absolute performance; detailed calculations are provided in Appendix N. As shown in Figure 4, Europe and North America tend to show stronger relative performance, whereas Africa, South America, and Oceania appear more frequently among weaker regions. The pattern is less consistent for image-only Interpretation but becomes clearer when metadata is provided. These results reveal systematic variation in model performance across cultural regions.

Web search enriches interpretation, but does not reliably improve factual faithfulness. Table 5 shows that web search consistently increases Completeness and Depth across all eight model–setting pairs, while Conciseness decreases in seven of them. Faithfulness improves in only three pairs and decreases for three of four models under metadata guidance. Identification is similarly mixed: the strict All score improves for three models, but Materials accuracy drops for all four. These results suggest that web search can enrich model outputs but does not make them uniformly more reliable, indicating that retrieved external information may introduce irrelevant or conflicting evidence, which can act as noise and interfere with the model's grounding in the target object.

5 Conclusion

In this work, we present MuseFACET to systematically evaluate the trustworthiness and controllability of MLLMs in museum collection interpretation. Through hierarchical Identification and Interpretation tasks, our evaluation reveals challenges in accurate object identification, factual grounding, and object-specific interpretation. The results further highlight the importance of reliable metadata in improving generation quality, while performance differences across cultural regions indicate the need for broader cultural coverage and more balanced model capabilities. These findings suggest that effective museum collection interpretation requires not only visual understanding but also accurate knowledge grounding and context-aware generation. We hope MuseFACET will support future research on reliable and culturally inclusive open-ended vision-to-language generation.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, including improving grammar, clarity, and readability. Generative AI tools were additionally used in a limited capacity for code assistance, primarily for debugging and syntactic correction; all AI-assisted code was reviewed and tested by the authors.

Generative AI tools were not used for research conceptualization, benchmark and evaluation design, dataset construction, or analysis and interpretation of experimental results. LLMs were used as automatic evaluators as part of the experimental methodology; this use is described in Section 4.1 and Appendix I. We reviewed and verified all AI-assisted work and take full responsibility for the scientific content and conclusions of this paper.

ETHICS STATEMENT

Human subjects. The construction and evaluation of MuseFACET involve human annotators for metadata verification and evaluation of model outputs. We provide detailed documentation of the annotation criteria, instructions, and evaluation procedures in Appendices E and M to support transparency and reproducibility.

Cultural representation and scope. While evaluating cultural understanding is a central goal of this work, we acknowledge that culture is inherently complex, dynamic, and context-dependent. Consequently, MuseFACET cannot capture the full richness of every cultural context. The benchmark may inherit representational biases from the source museum collections and catalog records, and some cultural regions, historical traditions, or object types may remain underrepresented. Museum records may also reflect the curatorial perspectives and cataloging conventions of their source institutions (Turner, 2020). To support consistent organization and cross-regional analysis across these heterogeneous sources, MuseFACET uses standardized geographic labels. These geographic labels provide a useful basis for comparison, but are not an exhaustive representation of cultural identity.

Interpretive validity and responsible use. Museum interpretation depends on culturally and historically specific concepts, terminology, and conventions. Model-generated descriptions may introduce conceptually misaligned claims, unsupported factual assertions, or overconfident attributions. Such errors may lead to misattributed traditions or distorted historical and cultural interpretations rather than merely lower benchmark scores. We therefore position MuseFACET as a diagnostic benchmark for evaluation and failure analysis, and model outputs should not be treated as authoritative museum interpretations or substitutes for expert judgment. Accordingly, our evaluation is designed to assess whether model-generated interpretations are both trustworthy and controllable.

REPRODUCIBILITY STATEMENT

We give extensive details of our setup in the Appendix. For dataset construction, we provide the data sources and acquisition procedures in Appendix A, the filtering and sampling procedures in Appendix D, and the human annotation protocol in Appendix E. For evaluation, we provide the complete task definitions and evaluation prompts in Appendix I, together with model and inference details in Appendix H. For human evaluation, we provide the annotation instructions and evaluation protocol in Appendix M. Several models used for evaluation are proprietary and accessible through APIs, so we cannot guarantee that future calls will produce exactly the same outputs. We plan to release the benchmark data permitted by the original museum licenses, together with the code used for data processing and evaluation, to facilitate replication of our work.

REFERENCES

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11569–11579, 2021.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language

model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18228–18258, 2025.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pp. 382–398. Springer, 2016.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Murtha Baca. *Cataloging cultural objects: A guide to describing cultural works and their images*. American Library Association, 2006.

Murtha Baca and Patricia Harpring. Categories for the description of works of art. 2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zechen Bai, Yuta Nakashima, and Noa Garcia. Explain me the painting: Multi-topic knowledgeable art description generation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5402–5412. IEEE, 2021.

Ada-Astrid Balauca, Sanjana Garai, Stefan Balauca, Rasesh Udayakumar Shetty, Naitik Agrawal, Dhwanil Subhashbhai Shah, Yuqian Fu, Xi Wang, Kristina Toutanova, Danda Pani Paudel, et al. Understanding museum exhibits using vision-language reasoning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2227–2238. IEEE, 2025.

J Martin Bland and DouglasG Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NVLM: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024.

Martin Doerr. The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75, 2003.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d79a27cf2772fe00be7f341efc0eb517-Paper-Datasets_and_Benchmarks_Track.pdf.

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

Sara Ghaboura, Ketan Pravin More, Ritesh Thawkar, Wafa Al Ghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. Time travel: A comprehensive benchmark to evaluate lmms on historical and cultural artifacts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23627–23641, 2025.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. Towards artwork explanation in large-scale vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 705–729, 2024.

Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2474–2495, 2024.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, 2024.

Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Liang, Yang Deng, et al. RAVENEA: A benchmark for multimodal retrieval-augmented visual culture understanding. In *International Conference on Learning Representations*, volume 2026, pp. 18290–18317, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*, 2024.

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3190–3199. IEEE, 2019.

Thomas Mensink and Jan Van Gemert. The rijksmuseum challenge: Museum-centered visual recognition. In *Proceedings of international conference on multimedia retrieval*, pp. 451–454, 2014.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5769–5790, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026a.

GLMV Team, Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanli Wang, et al. GLM-5V-Turbo: Toward a native foundation model for multimodal agents. *arXiv preprint arxiv:2604.26752*, 2026b.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Hannah Turner. Cataloguing culture. *Legacies of Colonialism in Museum Documentation, Vancouver*, 2020.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xuefeng Wei, Zhixuan Wang, Xuan Zhou, Zhi Qu, Hongyao Li, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. Cartbench: Evaluating vision-language models on chinese art understanding, interpretation, and authenticity. *arXiv preprint arXiv:2604.11632*, 2026.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, et al. Efficient gpt-4v level multimodal large language model for deployment on edge devices. *Nature Communications*, 16(1):5509, 2025.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets. In *International Conference on Learning Representations*, volume 2024, pp. 55361–55414, 2024.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*, 2024.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. MiniCPM-V 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6713–6724. IEEE, 2019.

Nonghai Zhang, Zeyu Zhang, Jiazi Wang, Yang Zhao, and Hao Tang. VaseVQA-3D: Benchmarking 3d VLMs on ancient greek pottery. In *International Conference on Learning Representations*, volume 2026, pp. 89839–89863, 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

Table of Contents

A Museum Sources, Data Acquisition, and Redistribution **17**

B Cross-Museum Schema Alignment **17**

C Metadata Normalization and Rule-Based Pre-annotation **18**

D Filtering and Diversity-Oriented Sampling **21**

E Human Annotation and Quality Control **22**

F Detailed Dataset Statistics and Representative Examples **29**

F.1 Overall Statistics and MuseFACET-testmini 30

F.2 Source-Museum and Geographic Coverage 31

F.3 Statistics by Object Category 31

F.4 Temporal and Textual Statistics 32

F.5 Representative Examples 32

G Task Design and Generation Prompts **46**

H Experimental Details and Model Configurations **47**

H.1 Proprietary MLLMs in the Main Evaluation 48

H.2 Open-Source MLLMs in the Main Evaluation 48

H.3 Evaluation Model 49

H.4 Additional Checkpoints in the Judge-Bias Analysis 49

H.5 Evaluated Models 50

H.6 Inference and Input Settings 50

H.7 Output Handling 50

I Evaluation Framework and Judge Prompts **51**

I.1 Identification Evaluation 51

I.2 Interpretation Evaluation 52

I.3 Date Evaluation 57

J Judge Self-Evaluation Bias Analysis **57**

K Qualitative Error Analysis **60**

K.1 Cross-cultural prototype substitution and task-dependent narrative drift 61

K.2 Event-state hallucination and metadata-induced narrative overcommitment 62

L Web-Search Evaluation **64**

L.1 Web-Search Setup 64

L.2 Complete No-Search Interpretation Results 64

L.3 Effect of Web Search 65

M Human Evaluation and LLM-Judge Validation **65**

M.1 Evaluation Sample 66

M.2 Annotators and Interface 66

M.3 Agreement Analysis 66

N Cultural-Region Analysis **67**

Table 7: Source museums used in MuseFACET. The reuse basis summarizes the official rights or terms applicable to the content selected for the benchmark. Links point to the corresponding official access or rights documentation.

Museum	Official access point	Reuse basis	Final
Cleveland Museum of Art	Open Access API	CC0 Open Access data and images	1,220
Minneapolis Institute of Art	Collection repository	CC0 collection metadata; Public Domain images	1,135
Art Institute of Chicago	Collection API / IIIF	CC0 API metadata; descriptions under CC BY 4.0; Public Domain images	835
Tokyo National Museum	ColBase API	Images and commentary reusable for commercial and non-commercial purposes with source attribution	657
Rijksmuseum	Rijksmuseum Data Services	Record-level PDM, CC0, or CC BY 4.0	500
Museum of New Zealand Te Papa Tongarewa	Collections API	CC BY 4.0 metadata; image reuse according to record-level rights and institutional permission	376
Smithsonian National Museum of Asian Art	Open Access API	Smithsonian Open Access data and media under CC0	204
Smithsonian National Museum of African Art	Open Access API	Smithsonian Open Access data and media under CC0	48
Kyushu National Museum	ColBase API	Images and commentary reusable for commercial and non-commercial purposes with source attribution	25
Total			**5,000**

A Museum Sources, Data Acquisition, and Redistribution

Data source. MuseFACET is constructed from the publicly accessible digital collections of nine museums. For each source institution, we obtain collection records through its official API, data service, or institutional repository. The collected records include object metadata, museum-provided descriptions, and images or image references. Table 7 summarizes the official access points, applicable reuse basis, and the number of objects retained in the final benchmark.

Rights and redistribution. The source institutions adopt different rights and reuse frameworks. Before inclusion in the released benchmark, we verify that the selected content can be redistributed under the applicable licenses, rights statements, terms of use, or institutional permission. MuseFACET releases the corresponding object images, normalized metadata, and museum-provided descriptions. For each object, we retain the source institution and original record identifier to preserve provenance and enable traceability to the source collection.

B Cross-Museum Schema Alignment

Museum collection systems organize equivalent information using different cataloging conventions and field structures. To integrate records from the nine source museums into a common representation, we align museum-specific fields to eight shared information fields used in MuseFACET: *category*, *title*, *date*, *country*, *materials*, *artist*, *description*, and image information. This alignment establishes a common structural representation before any field values are normalized.

The mapping is not necessarily one-to-one. Museum documentation practices may represent the same semantic concept through multiple fields serving different display, indexing, or contextual functions, as reflected in professional cataloging frameworks such as CCO and CDWA (Baca, 2006; Baca & Harpring, 2017). We therefore identify all source fields that provide relevant evidence for each shared field rather than selecting a single source value at this stage. Tables 8–15 report these mappings for each field. Tokyo National Museum and Kyushu National Museum share the ColBase schema, while the two Smithsonian museums share the Smithsonian Open Access schema.

Table 8: Source fields aligned to the category field.

Source	Source field(s)
Art Institute of Chicago	`artwork_type_title` `classification_title`
Cleveland Museum of Art	`type` `classification`
Minneapolis Institute of Art	`classification` `object_name`
ColBase (Tokyo / Kyushu)	`bunrui` `categories`
Te Papa	`isTypeOf` `additionalType`
Smithsonian NMAA / NMAfA	`content.freetext.objectType` `content.indexedStructured.object_type`
Rijksmuseum	`object_name`

Table 9: Source fields aligned to the title field.

Source	Source field(s)
Art Institute of Chicago	`title`
Cleveland Museum of Art	`title`
Minneapolis Institute of Art	`title`
ColBase (Tokyo / Kyushu)	`title`
Te Papa	`title`
Smithsonian NMAA / NMAfA	`content.descriptiveNonRepeating.title`
Rijksmuseum	`title`

Schema alignment preserves the relevant information available in the original museum records without determining the final normalized value. Decisions about field precedence, value extraction, parsing, terminology, and standardization are made in the subsequent field-specific normalization stage. The complete normalization procedures are described in Appendix C.

C Metadata Normalization and Rule-Based Pre-annotation

Following schema alignment, we further normalize selected metadata fields to reduce differences in granularity, terminology, and format across museums. The normalization focuses on *category*, *country*, *date*, and *materials*, which exhibit the largest variation across source collections. We apply museum-specific mappings for category and country, and rule-based parsing and extraction for date and materials. The resulting values serve as pre-annotations and are subsequently checked during human annotation.

Category. Museum collections use different classification systems and often describe objects at different levels of granularity. We map the source object types and classifications to ten shared categories: Painting, Drawing, Calligraphy, Sculpture, Ceramic Object, Metalwork, Glass Object, Furniture, Arms and Armor, and Textile and Costume. Museum-specific mapping tables are used to account for differences in terminology across collections. Fine-grained source labels are mapped to a broader category when their meanings are compatible, while ambiguous or potentially misleading mappings are reserved for manual review.

Country. Geographic information is recorded in heterogeneous forms, including countries, cities, regions, historical places, and cultural attributions. We normalize this information to a country-

Table 10: Source fields aligned to the date field.

Source	Source field(s)

Art Institute of Chicago	`date_display`
	`date_start`
	`date_end`

Cleveland Museum of Art	`creation_date`
	`creation_date_earliest`
	`creation_date_latest`

Minneapolis Institute of Art	`dated` `date`

ColBase (Tokyo / Kyushu)	`jidai_seiki`
	`jidai_from`
	`jidai_to`

Te Papa	`production[].verbatimCreatedDate`
	`production[].createdDate`
	`production[].facetCreatedDate`

Smithsonian NMAA / NMAfA	`content.freetext.date[]`

Rijksmuseum	`date_begin`
	`date_end`
	`date_begin_bce`
	`date_end_bce`

Table 11: Source fields aligned to country and geographic-origin information.

Source	Source field(s)
Art Institute of Chicago	`place_of_origin` `style_title`
Cleveland Museum of Art	`culture` `find_spot`

Minneapolis Institute of Art	`country`
	`culture`
	`continent`

ColBase (Tokyo / Kyushu)	`seisakuchi`
	`shutudochi`
	`jidai_seiki`

Te Papa	`production[].spatial.title`
	`production[].spatial.prefLabel`
	`production[].spatial.nation`
	`production[].contributor.nationality`

Smithsonian NMAA / NMAfA	`content.freetext.place[]`
	`content.indexedStructured.place`
	`content.freetext.culture[]`

Rijksmuseum	`location` `culture`

level label when the museum record provides sufficient evidence for the object's place of creation or origin(Baca & Harpring, 2017; Baca, 2006). Explicit production- or origin-place information is preferred when available, while artist nationality, museum location, depicted place, and findspot are not used as direct substitutes for object origin. Museum-specific place mappings are used to standardize equivalent geographic expressions. When a historical or regional attribution cannot be reliably reduced to a single modern country, we retain the source-supported label rather than forcing a potentially incorrect mapping.

Table 12: Source fields aligned to the materials field.

Source	Source field(s)
Art Institute of Chicago	material_titles medium_display technique_titles
Cleveland Museum of Art	technique medium
Minneapolis Institute of Art	medium
ColBase (Tokyo / Kyushu)	hinshitu_keijo
Te Papa	isMadeOf isMadeOfSummary productionUsedTechnique
Smithsonian NMAA / NMAfA	content.freetext.physicalDescription[] content.indexedStructured.material
Rijksmuseum	materials

Table 13: Source fields aligned to the artist field.

Source	Source field(s)
Art Institute of Chicago	artist_title artist_display
Cleveland Museum of Art	creators[].name creators[].role creators[].description
Minneapolis Institute of Art	artist maker role
ColBase (Tokyo / Kyushu)	sakusha
Te Papa	production[].contributor.title production[].role
Smithsonian NMAA / NMAfA	content.freetext.name[] content.indexedStructured.name
Rijksmuseum	artist_name artist_role artist_nationality

Date. Museum dates appear in a wide range of formats, including exact years, year ranges, decades, centuries, approximate dates, and BCE expressions. We convert these expressions into a single year or a year interval while preserving the temporal precision of the original record(Baca & Harpring, 2017; Baca, 2006). For example, "18th century" is represented as the interval "1701–1800" rather than as a single representative year. When both a human-readable display date and structured numerical bounds are available, the display date is used to preserve the intended interpretation, while the numerical fields provide supporting information. BCE dates and ranges spanning BCE and CE are handled explicitly.

Materials. Museums differ substantially in how materials and production techniques are recorded. Some collections provide dedicated material fields, while others combine materials, supports, and techniques in a single medium expression(Baca & Harpring, 2017; Baca, 2006). We extract the principal physical materials and represent them as a standardized list, while excluding technique-only terms. For example, "oil on canvas" is normalized to retain both the paint medium and the support, such as "oil paint; canvas". We also remove straightforward duplicates and consolidate equivalent

Table 14: Source fields aligned to the description field.

Source	Source field(s)
Art Institute of Chicago	`description` `short_description`
Cleveland Museum of Art	`description`
Minneapolis Institute of Art	`text` `description`
ColBase (Tokyo / Kyushu)	`descriptions[].text`

Te Papa	`description`
	`caption`
	`captionFormatted`

Smithsonian NMAA / NMAfA	`content.freetext.summary[]` `content.freetext.notes[]`
Rijksmuseum	`description`

Table 15: Source fields aligned to image information.

Source	Source field(s)
Art Institute of Chicago	`image_id` `_iiif_url`

Cleveland Museum of Art	`images.web.url`
	`images.print.url`
	`images.full.url`

Minneapolis Institute of Art	`Cache_Location` `Primary_RenditionNumber`
ColBase (Tokyo / Kyushu)	`image_files[].url` `image_files[].slideshow_l_url`

Te Papa	`hasRepresentation[].iiifUrl`
	`hasRepresentation[].contentUrl`
	`hasRepresentation[].previewUrl`
	`hasRepresentation[].thumbnailUrl`

Smithsonian NMAA / NMAfA	`content.descriptiveNonRepeating.`
	`online_media.media[].content`
	`content.descriptiveNonRepeating.`
	`online_media.media[].thumbnail`
	`content.descriptiveNonRepeating.`
	`online_media.media[].resources[].url`

Rijksmuseum	`image_url`

expressions when they refer to the same material, while retaining distinct materials supported by the source record.

These rules are intended to provide consistent pre-annotations rather than to replace item-level judgment. Museum records may contain ambiguous terminology, source-specific conventions, or exceptional cases that cannot be handled reliably by general rules. The sampled records are therefore subsequently reviewed against the original museum evidence, and necessary corrections are made during the human annotation and quality-control stage described in Appendix E.

D Filtering and Diversity-Oriented Sampling

Following schema alignment and rule-based pre-annotation, we further filter the collected records and perform diversity-oriented sampling to construct the benchmark. The filtering stage removes

records that cannot reliably support both Identification and Interpretation, while the sampling stage improves coverage across object categories and geographic origins without forcing an artificially uniform distribution.

Initial filtering. We first remove records that do not provide the information required by our tasks. Each retained record must contain an accessible object image, the required task-relevant metadata, and a substantive museum-provided description. Records with missing or unusable images, missing required metadata, or descriptions that are missing or too short are excluded. Artist information is not required because many museum objects are anonymous or unattributed. We also remove records for which the available image or description does not provide sufficiently reliable object-level information for subsequent evaluation.

Candidate pool. After initial filtering and pre-annotation, we obtain a candidate pool in which each record is associated with a normalized category and country label. These labels allow us to organize the candidates according to their joint category–country distribution. Because the source collections are naturally imbalanced, frequent categories and countries can otherwise dominate a random sample, while less frequent category–country combinations receive limited representation.

Diversity-oriented sampling. We perform stratified sampling over the joint distribution of category and country, treating each category–country combination as a stratum. Rather than allocating samples in direct proportion to stratum size, we use *square-root allocation*, so that the sampling quota of a stratum grows approximately with the square root of its number of eligible records. This reduces the dominance of frequent strata while assigning relatively greater sampling weight to less frequent combinations, without enforcing equal sample sizes across strata. We further apply a greedy selection procedure that prioritizes previously uncovered category–country combinations, thereby increasing joint coverage under the fixed sampling budget. When comparable candidates are available, records with more substantive museum-provided descriptions are preferred. Together, these procedures retain the natural variation of the source collections while improving representation across object categories and geographic origins.

Post-sampling verification. The sampled records are subsequently reviewed through the human annotation and quality-control procedure described in Appendix E. Annotators verify the normalized metadata, image–record correspondence, and description quality against the original museum evidence. Records with unresolved metadata ambiguity, mismatched images, unsuitable descriptions, or other substantive quality problems are removed. When removals reduce the dataset below the target size, additional records are sampled from the candidate pool and undergo the same annotation and quality-control procedure before inclusion in the final benchmark.

E Human Annotation and Quality Control

Annotation protocol. Nine annotators conduct the formal audit using a shared web-based interface and a common set of annotation guidelines. The initial 5,000 sampled records are randomly assigned across annotators, with each record reviewed by one primary annotator. Two annotators review 750 records each, while the remaining seven review 500 records each.

Before formal annotation, the three researchers developing the guidelines review the same pilot set of 300 records. The pilot is used to identify ambiguous cases, compare alternative normalization decisions, and refine the operational guidelines before the full audit. For example, museum sources may provide both human-readable textual dates and structured numeric date bounds. For the Art Institute of Chicago, the official API documentation [2] identifies `date_display` as the preferred field for display on websites and applications, while `date_start` and `date_end` provide numeric bounds for the creation period. We therefore use the museum's displayed date information as the primary evidence when interpreting the catalog date, with structured bounds serving as supporting information.

During formal annotation, the interface presents the object image, normalized metadata, museum-provided description, and the corresponding raw museum values. This allows annotators to compare

[2] `https://api.artic.edu/docs/`

each normalized field directly against the source evidence rather than judging the normalized record in isolation.Figures 5–10 present the complete annotation guidelines used in the manual verification process

Museum Object Data Annotation Guideline

1. Background and annotation task

This study examines how well multimodal models can identify and describe museum objects. To construct the evaluation data, we collect object records from several museums and convert their metadata into a common format. Because museums use different cataloguing conventions, and automatic processing may introduce errors, the selected records need to be checked manually before they are used for evaluation.

For each record, you will see the object image, the current normalized metadata, the museum-provided description, and the corresponding source information. Your task is to check whether the current record is consistent with the original museum evidence and suitable for use in the dataset.

The review is issue-based. If you find a problem, select the corresponding issue type. If no issue is found after checking the record, leave all issue boxes unchecked. More than one issue may be selected for the same record.

> **Important:** Base your decision on the museum record and the displayed image. Do not add or correct information using outside knowledge when it is not supported by the source. If the available evidence is ambiguous or contradictory, flag the relevant field and briefly explain the uncertainty.

2. Annotation criteria

1. Image / record mismatch

Check whether the image, title, metadata, and description refer to the same catalogued object. An alternate view or detail of the same object is acceptable.

Flag this issue if the image shows a different object, the wrong member of a series, or otherwise does not correspond to the catalog record.

2. Category issue

The dataset uses ten object categories: **Painting, Drawing, Calligraphy, Sculpture, Ceramic Object, Metalwork, Glass Object, Furniture, Arms and Armor,** and **Textile and Costume**.

Check whether the source object type or classification has been mapped to an appropriate category. A detailed museum term may be mapped to a broader category if the meaning is preserved.

> Example: `Bust` may reasonably be mapped to `Sculpture`. A museum record explicitly classified as a print should not be treated as `Drawing` simply because both are works on paper.

3. Title issue

Check whether the current title refers to the same object or subject as the museum title and remains readable.

- Translations, proper names, bilingual titles, punctuation cleanup, and harmless abbreviation are acceptable.
- Flag broken text, leftover HTML or encoding artifacts, catalog-number-only titles, or a title that changes the identity or subject of the object.

4. Materials issue

Check whether the normalized materials preserve the principal physical substances supported by the museum record.

- Do not omit a core material or support.
- Do not add a material that is not supported by the source.
- Do not treat techniques such as weaving, casting, embroidery, or appliqué as materials.
- Avoid redundant parent/child or singular/plural duplicates.

> Example: `oil on canvas` should retain both the paint medium and the support, such as `oil paint; canvas`.

5. Date issue

Check whether the normalized date correctly represents the object's museum date as a year or year range.

- Preserve the intended range when the source gives a range.
- Interpret century and decade expressions carefully.
- Do not reverse BCE chronology.

Figure 5: Annotation guidelines used for manual verification (page 1 of 6).

- Do not mix the object's creation date with restoration, acquisition, or scientific-test dates.

When a museum provides both a human-readable date and numerical bounds, read the displayed date first and use the bounds as supporting evidence.

6. Country issue

Check whether the current country is supported as the object's place of creation or origin. Explicit production or origin information should be preferred when available.

Do not use the artist's nationality, later residence, findspot, depicted place, or the museum's location as the object's country of origin. Historical geographic labels may be retained when converting them to a modern country would change or overstate the museum evidence.

7. Artist issue

Check whether the current artist or maker agrees with the museum creator information. A blank value is acceptable when the museum record does not support a named creator.

- Preserve meaningful uncertainty such as *attributed to* when relevant.
- For collaborative works, check whether important named contributors have been omitted or misidentified.

8. Description issue

The description should be readable, substantive, and related to the specific object. It may include visual features, subject matter, construction, function, historical context, or museum interpretation.

Flag a description that belongs to another object, is unreadable, contains only catalog metadata, or gives only generic series-level information without describing the individual object.

Other issue

Use **Other issue** only for a substantive problem that is not covered by the eight checks above. Briefly explain the reason in the note field.

3. General decision rules

Situation	How to handle it
Multiple source fields	Use all relevant museum evidence. Different source fields may provide complementary information about the same normalized field.
Harmless wording difference	Do not flag a difference if the meaning is preserved after normalization.
Ambiguous evidence	Flag the relevant field and explain the ambiguity instead of forcing a correction.
Image evidence	The image can help check object consistency and category, but it should not be used to invent unsupported metadata.
Unresolvable record	Flag the record for later review. Do not guess a final value.

4. Short examples

> **Materials:** The museum record lists the medium as `Oil on canvas`, while the current value contains only `canvas`. Flag **Materials issue** because a core material has been omitted.

> **Date:** The museum record gives the date as `c. 1618–20`, while the current value is `1618`. Flag **Date issue** because the documented interval has been truncated.

> **Country:** The museum record gives a production place and a separate findspot. Use the production-place evidence for origin; do not replace it with the findspot.

5. Source-field reference

The tables below reproduce the source-field mappings reported in Appendix B. They show which museum-specific fields are aligned to each shared field used in the review.

Figure 6: Annotation guidelines used for manual verification (page 2 of 6).

Category

Source	Source field(s)
Art Institute of Chicago	`artwork_type_title` `classification_title`
Cleveland Museum of Art	`type` `classification`
Minneapolis Institute of Art	`classification` `object_name`
ColBase (Tokyo / Kyushu)	`bunrui` `categories`
Te Papa	`isTypeOf` `additionalType`
Smithsonian NMAA / NMAfA	`content.freetext.objectType` `content.indexedStructured.object_type`
Rijksmuseum	`object_name`

Title

Source	Source field(s)
Art Institute of Chicago	`title`
Cleveland Museum of Art	`title`
Minneapolis Institute of Art	`title`
ColBase (Tokyo / Kyushu)	`title`
Te Papa	`title`
Smithsonian NMAA / NMAfA	`content.descriptiveNonRepeating.title`
Rijksmuseum	`title`

Figure 7: Annotation guidelines used for manual verification (page 3 of 6).

Date

Source	Source field(s)

Art Institute of Chicago	`date_display`
	`date_start`
	`date_end`

Cleveland Museum of Art	`creation_date`
	`creation_date_earliest`
	`creation_date_latest`

Minneapolis Institute of Art	`dated` `date`

ColBase (Tokyo / Kyushu)	`jidai_seiki`
	`jidai_from`
	`jidai_to`

Te Papa	`production[].verbatimCreatedDate`
	`production[].createdDate`
	`production[].facetCreatedDate`

Smithsonian NMAA / NMAfA	`content.freetext.date[]`

Rijksmuseum	`date_begin`
	`date_end`
	`date_begin_bce`
	`date_end_bce`

Country and geographic origin

Source	Source field(s)
Art Institute of Chicago	`place_of_origin` `style_title`
Cleveland Museum of Art	`culture` `find_spot`

Minneapolis Institute of Art	`country`
	`culture`
	`continent`

ColBase (Tokyo / Kyushu)	`seisakuchi`
	`shutudochi`
	`jidai_seiki`

Te Papa	`production[].spatial.title`
	`production[].spatial.prefLabel`
	`production[].spatial.nation`
	`production[].contributor.nationality`

Smithsonian NMAA / NMAfA	`content.freetext.place[]`
	`content.indexedStructured.place`
	`content.freetext.culture[]`

Rijksmuseum	`location` `culture`

Figure 8: Annotation guidelines used for manual verification (page 4 of 6).

Materials

Source	Source field(s)

Art Institute of Chicago	`material_titles`
	`medium_display`
	`technique_titles`

Cleveland Museum of Art	`technique` `medium`
Minneapolis Institute of Art	`medium`
ColBase (Tokyo / Kyushu)	`hinshitu_keijo`

Te Papa	`isMadeOf`
	`isMadeOfSummary`
	`productionUsedTechnique`

Smithsonian NMAA / NMAfA	`content.freetext.physicalDescription[]` `content.indexedStructured.material`
Rijksmuseum	`materials`

Artist

Source	Source field(s)
Art Institute of Chicago	`artist_title` `artist_display`

Cleveland Museum of Art	`creators[].name`
	`creators[].role`
	`creators[].description`

Minneapolis Institute of Art	`artist`
	`maker`
	`role`

ColBase (Tokyo / Kyushu)	`sakusha`
Te Papa	`production[].contributor.title` `production[].role`
Smithsonian NMAA / NMAfA	`content.freetext.name[]` `content.indexedStructured.name`

Rijksmuseum	`artist_name`
	`artist_role`
	`artist_nationality`

Figure 9: Annotation guidelines used for manual verification (page 5 of 6).

Description

Source	Source field(s)
Art Institute of Chicago	`description` `short_description`
Cleveland Museum of Art	`description`
Minneapolis Institute of Art	`text` `description`
ColBase (Tokyo / Kyushu)	`descriptions[].text`

Te Papa	`description`
	`caption`
	`captionFormatted`

Smithsonian NMAA / NMAfA	`content.freetext.summary[]` `content.freetext.notes[]`
Rijksmuseum	`description`

Image information

Source	Source field(s)
Art Institute of Chicago	`image_id` `_iiif_url`

Cleveland Museum of Art	`images.web.url`
	`images.print.url`
	`images.full.url`

Minneapolis Institute of Art	`Cache_Location` `Primary_RenditionNumber`
ColBase (Tokyo / Kyushu)	`image_files[].url` `image_files[].slideshow_l_url`

Te Papa	`hasRepresentation[].iiifUrl`
	`hasRepresentation[].contentUrl`
	`hasRepresentation[].previewUrl`
	`hasRepresentation[].thumbnailUrl`

Smithsonian NMAA / NMAfA	`content.descriptiveNonRepeating.online_media.media[].content`
	`content.descriptiveNonRepeating.online_media.media[].thumbnail`
	`content.descriptiveNonRepeating.online_media.media[].resources[].url`

Rijksmuseum	`image_url`

A flag indicates that the record needs follow-up review. It does not by itself change or remove the record.

Figure 10: Annotation guidelines used for manual verification (page 6 of 6).

Table 16: Issue checks used in the manual audit. Annotators flag a record when the normalized data fail the corresponding verification criterion. An additional "Other" option is available for cases not covered by the eight predefined checks.

Issue check	Verification criterion
Image–record mismatch	The image, title, and description should refer to the same cataloged object.
Category	The source object type should be mapped to the appropriate one of the ten MuseFACET categories.
Title	The normalized title should faithfully reflect the source title, remain readable in English, and be suitable as an Identification target.
Materials	Core materials should be retained, unsupported materials should not be introduced, and techniques should not be treated as materials.
Date	The normalized year or range should correctly reflect the source date, including qualifiers such as circa, century expressions, and BCE dates.
Country	The country assignment should be supported by museum evidence concerning place, production, origin, or culture, with the museum description consulted when necessary.
Artist	The normalized artist information should agree with the source creator or maker fields and should not omit relevant contributors.
Description	The museum-provided description should be readable, substantive, and actually concern the corresponding object.

Verification criteria. The audit uses eight predefined issue checks covering image–record consistency, category, title, materials, date, country, artist, and description. An additional "Other" option is provided for cases not captured by these categories. The operational criteria shown to annotators are summarized in Table 16.

Figure 11 shows the web-based interface used in the formal audit. For metadata fields requiring normalization, the raw museum values are displayed alongside the current normalized values so that annotators can evaluate the mapping directly.

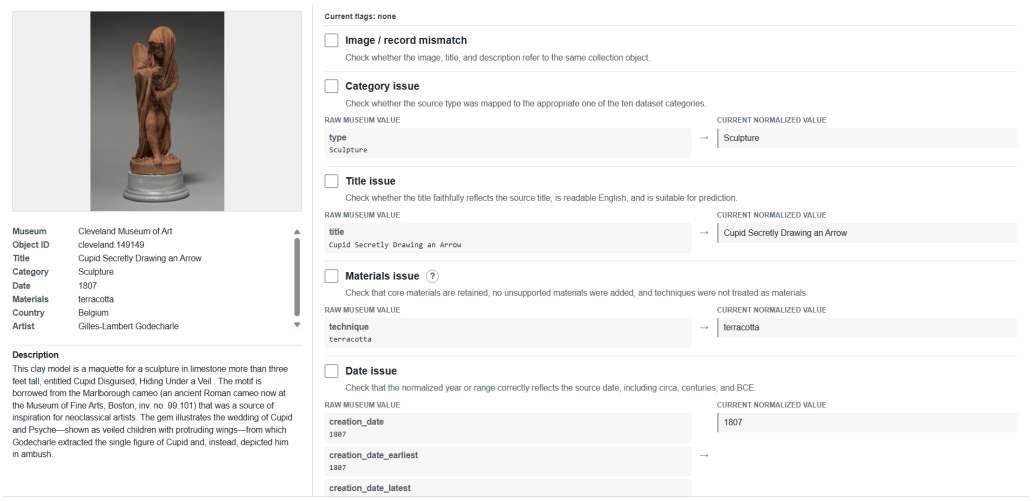
Audit findings. In the initial 5,000-record audit, 939 records (18.8%) are flagged with at least one issue. Because a record may receive multiple flags, the audit yields 1,034 issue flags in total. Table 17 reports their distribution. Materials account for nearly half of all detected issues, followed by date, description, and artist. Materials and date together account for 73.6% of all issue flags, indicating that fields requiring more extensive cross-museum normalization are also the most frequent sources of problems identified during manual review.

Quality control. The three guideline developers subsequently focus their quality-control review on the 939 flagged records. Records with an image–record mismatch or an unsuitable museum description are removed because they cannot reliably support the benchmark tasks. For the remaining metadata-related issues, reviewers verify the annotations and proposed corrections against the original museum evidence. Most cases can be resolved directly under the established guidelines, while a smaller number of complex or ambiguous cases are discussed jointly before a final decision is made.

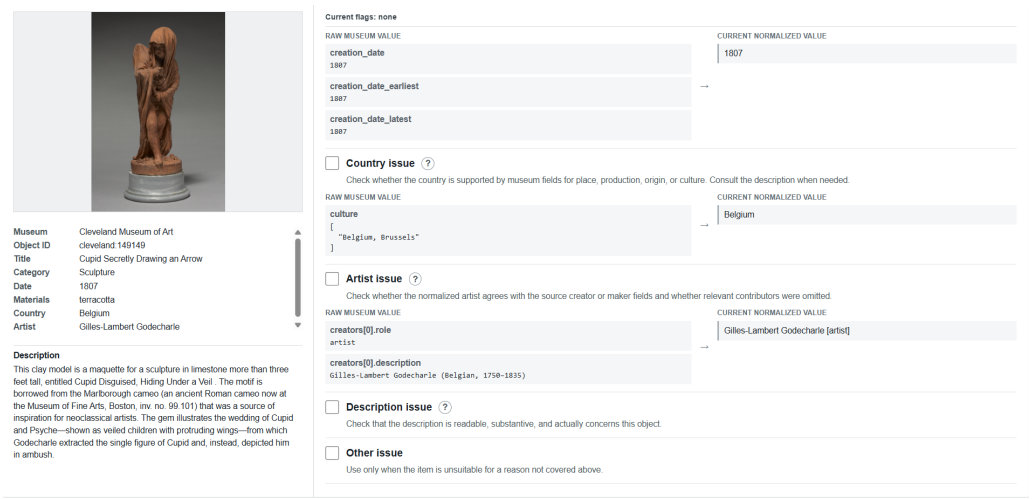
In total, 151 records are removed during this quality-control stage. Additional candidates are then sampled to restore the target dataset size. Every replacement record undergoes the same annotation and quality-control procedure before inclusion, and the process is repeated until the final 5,000-record benchmark satisfies the same verification criteria.

F Detailed Dataset Statistics and Representative Examples

This appendix provides detailed statistics for the full 5,000-object MuseFACET benchmark and the 1,000-object *MuseFACET-testmini* subset. In addition to the aggregate statistics reported in the main paper, we report distributions across source museums, object categories, geographic origins, temporal periods, and textual fields.



(a) Annotate interface (1)



(b) Annotate interface (2)

Figure 11: Interface used for manual verification. The left side presents the object image, normalized metadata, and museum-provided description, while the review panel presents the predefined issue checks. For metadata fields requiring normalization, the raw museum values are shown alongside the current normalized values for source-grounded verification. Panels (a) and (b) show different portions of the same review interface.

F.1 Overall Statistics and MuseFACET-testmini

Table 18 summarizes the overall composition of the full benchmark and MuseFACET-testmini. The full dataset contains objects from nine museums, ten canonical categories, and 89 country-level origins, covering 316 distinct category–country combinations. MuseFACET-testmini retains all nine museums and ten categories, together with 61 country-level origins and 191 category–country combinations.

The subset closely preserves the major marginal distributions of the full benchmark. In particular, the proportion of each object category differs by at most 0.06 percentage points between the two datasets. The distributions across source museums and continent-level cultural regions are also closely aligned, while the reduced subset naturally contains fewer rare country-level origins and category–country combinations.

Table 17: Distribution of issue flags identified during the initial manual audit. A record may receive multiple flags. In total, 939 of 5,000 records (18.8%) receive at least one flag.

Issue type	Count	Share
Materials	507	49.0%
Date	254	24.6%
Description	109	10.5%
Artist	97	9.4%
Country	26	2.5%
Image–record mismatch	18	1.7%
Other	15	1.5%
Category	5	0.5%
Title	3	0.3%
Total	1,034	100.0%

Table 18: Overall statistics of MuseFACET and MuseFACET-testmini.

Statistic	MuseFACET	MuseFACET-testmini
Objects	5,000	1,000
Source museums	9	9
Canonical categories	10	10
Country-level origins	89	61
Category–country combinations	316	191
Museum–category combinations	68	62
Museum–country combinations	257	143
Unique normalized material terms	325	174
Temporal span	4000 BCE–2022 CE	2454 BCE–2011 CE
Required benchmark fields complete	5,000 (100%)	1,000 (100%)
Artist information available	4,268 (85.4%)	854 (85.4%)

F.2 Source-Museum and Geographic Coverage

Table 19 reports the contribution of each source museum. In addition to object counts, we report the number of categories and country-level origins represented by each museum. The full benchmark draws most heavily from the Cleveland Museum of Art, Minneapolis Institute of Art, and Art Institute of Chicago, while preserving contributions from all nine sources in MuseFACET-testmini.

Country-level origins are grouped into the same six continent-level cultural regions used in the cultural analysis. As shown in Table 20, Europe and Asia constitute the two largest regions, while Africa, North America, Oceania, and South America remain represented in both the full benchmark and testmini.

Table 21 further reports the geographic composition within each source museum.

For completeness, Table 22 reports all 89 normalized country-level origins in the full benchmark together with their corresponding counts in MuseFACET-testmini. Countries are ordered by their frequency in the full benchmark.

F.3 Statistics by Object Category

The ten canonical categories differ substantially in both frequency and geographic coverage. Table 23 reports their sizes and the number of museums and country-level origins represented within each category. MuseFACET-testmini closely preserves the category distribution of the full benchmark.

Table 24 replaces a dense heatmap representation with the exact category-by-region counts.

Table 25 gives the corresponding source-museum by category matrix. The distribution shows that category coverage is not tied to a single institution: most categories are represented across multiple museums, while the source collections retain their naturally different collection profiles.

Table 19: Distribution by source museum. "F/M" reports the corresponding coverage in the full MuseFACET benchmark and MuseFACET-testmini, respectively.

Museum	Full	%	Mini	%	Categories F/M	Countries F/M
Cleveland Museum of Art	1,220	24.4	248	24.8	10/10	65/40
Minneapolis Institute of Art	1,135	22.7	226	22.6	10/10	72/41
Art Institute of Chicago	835	16.7	168	16.8	9/9	45/22
Tokyo National Museum	657	13.1	131	13.1	7/7	12/6
Rijksmuseum	500	10.0	99	9.9	9/9	27/13
Museum of New Zealand Te Papa Tongarewa	376	7.5	74	7.4	9/8	13/8
Smithsonian National Museum of Asian Art	204	4.1	40	4.0	5/4	8/5
Smithsonian National Museum of African Art	48	1.0	10	1.0	4/2	12/7
Kyushu National Museum	25	0.5	4	0.4	5/3	3/1
Total	**5,000**	**100.0**	**1,000**	**100.0**	–	–

Table 20: Distribution across continent-level cultural regions.

Region	Full	%	Mini	%
Africa	491	9.8	97	9.7
Asia	1,600	32.0	320	32.0
Europe	1,916	38.3	384	38.4
North America	560	11.2	113	11.3
Oceania	258	5.2	52	5.2
South America	175	3.5	34	3.4
Total	**5,000**	**100.0**	**1,000**	**100.0**

F.4 Temporal and Textual Statistics

Figure 12 provides the detailed temporal and description-length distributions. MuseFACET spans from approximately 4000 BCE to 2022 CE, with substantial representation from both historical and modern periods. Museum-provided descriptions also vary considerably in length, providing both relatively concise and substantially more detailed interpretive references.

Table 26 provides the corresponding temporal statistics for both dataset scales. Approximately half of the objects are represented by a single normalized year and half by a date interval. The three broad temporal bins remain closely aligned between MuseFACET and MuseFACET-testmini.

Table 27 summarizes the lengths of the principal textual and structured fields used by the benchmark. The testmini statistics remain close to those of the full benchmark across title length, number of materials, and description length.

For the full benchmark, 2,513 descriptions (50.3%) contain fewer than 100 words, 1,843 (36.9%) contain 100–199 words, and 644 (12.9%) contain at least 200 words. The corresponding counts in MuseFACET-testmini are 490 (49.0%), 372 (37.2%), and 138 (13.8%), respectively.

F.5 Representative Examples

Figures 13–22 present one representative record from each of the ten canonical object categories in MuseFACET. Each example shows the object image together with its normalized task-relevant metadata and museum-provided description. These examples illustrate the diversity of visual appearance, physical medium, geographic origin, historical period, and descriptive content represented in the benchmark.

Table 21: Number of MuseFACET objects by source museum and continent-level cultural region.

Museum	Africa	Asia	Europe	North America	Oceania	South America	Total
Cleveland Museum of Art	219	272	518	96	19	96	1,220
Minneapolis Institute of Art	119	281	441	232	28	34	1,135
Art Institute of Chicago	105	64	453	170	0	43	835
Tokyo National Museum	0	656	0	0	1	0	657
Rijksmuseum	0	148	348	1	1	2	500
Museum of New Zealand Te Papa Tongarewa	0	5	155	7	209	0	376
Smithsonian National Museum of Asian Art	0	149	1	54	0	0	204
Smithsonian National Museum of African Art	48	0	0	0	0	0	48
Kyushu National Museum	0	25	0	0	0	0	25
Total	**491**	**1,600**	**1,916**	**560**	**258**	**175**	**5,000**

Table 22: Country-level origin distribution in MuseFACET and MuseFACET-testmini.

Country	Full	Mini	Country	Full	Mini	Country	Full	Mini
China	601	121	Japan	600	122	United States	508	103
France	444	89	Italy	335	67	United Kingdom	322	65
Netherlands	283	58	Egypt	242	50	New Zealand	211	44
Germany	178	40	India	149	31	Peru	139	29
Belgium	118	23	Spain	57	13	Austria	48	9
Iran	47	9	Nigeria	46	11	Greece	45	9
Morocco	35	7	Indonesia	33	6	Philippines	31	6
Democratic Republic of the Congo	30	5	Ethiopia	28	4	Papua New Guinea	27	5
Mexico	24	5	Ghana	23	4	Colombia	21	4
Switzerland	18	3	Denmark	15	2	Mali	15	3
Cambodia	14	2	Korea	14	4	Algeria	13	2
Czech Republic	13	2	Nepal	12	3	Thailand	12	3
Turkey	11	3	Syria	11	1	South Africa	11	2
Pakistan	11	2	Panama	11	2	Russia	11	1
Guatemala	10	3	Sweden	10	1	Cameroon	10	3
Bolivia	9	1	Vietnam	9	1	Australia	9	1
Norway	8	1	Iraq	8	3	Ireland	6	1
Afghanistan	6	1	Sri Lanka	6	0	Myanmar	6	1
Sudan	5	1	French Polynesia	5	2	Laos	5	1
Burkina Faso	4	1	Canada	4	0	Uzbekistan	4	0
Uganda	4	1	Madagascar	4	1	Hungary	3	0
Kenya	3	0	Gabon	3	0	Fiji	3	0
Tunisia	3	0	Niger	3	1	Malaysia	3	0
Cyprus	2	0	Senegal	2	0	Zimbabwe	2	0
Chile	2	0	Cook Islands	2	0	Costa Rica	2	0
Bhutan	2	0	Suriname	2	0	Congo	2	1
Sierra Leone	2	0	Finland	1	0	Angola	1	0
Argentina	1	0	North Korea	1	0	Tonga	1	0
Poland	1	0	Bangladesh	1	0	Ecuador	1	0
Kyrgyzstan	1	0	Curaçao	1	0			

Table 23: Distribution and coverage by canonical object category. "F/M" reports coverage in MuseFACET and MuseFACET-testmini, respectively.

Category	Full	%	Mini	%	Museums F/M	Countries F/M
Painting	1,126	22.5	225	22.5	9/9	32/24
Drawing	849	17.0	170	17.0	6/6	24/14
Sculpture	698	14.0	140	14.0	8/8	56/37
Textile and Costume	670	13.4	134	13.4	8/6	52/34
Ceramic Object	558	11.2	111	11.1	8/6	39/21
Metalwork	385	7.7	77	7.7	6/5	43/25
Calligraphy	288	5.8	58	5.8	4/4	5/3
Furniture	214	4.3	43	4.3	6/5	28/13
Arms and Armor	150	3.0	30	3.0	8/8	25/13
Glass Object	62	1.2	12	1.2	5/5	12/7
Total	**5,000**	**100.0**	**1,000**	**100.0**	–	–

Table 24: Number of MuseFACET objects by canonical category and continent-level cultural region.

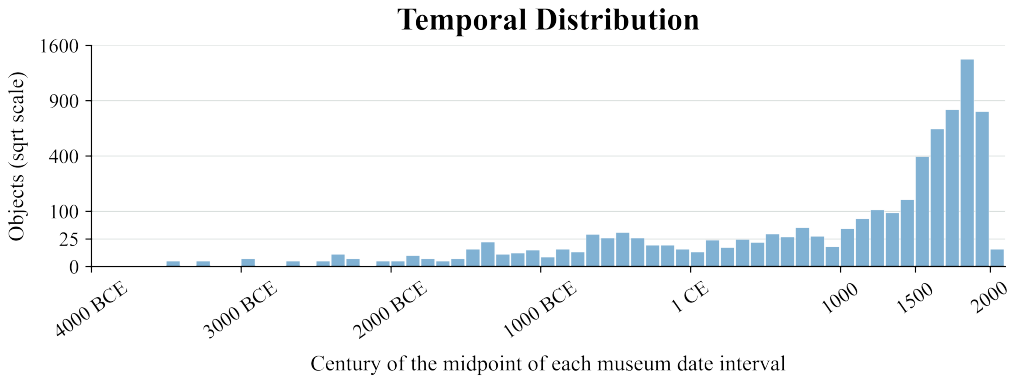
Category	Africa	Asia	Europe	North America	Oceania	South America	Total
Painting	27	547	454	73	24	1	1,126
Drawing	1	30	572	146	99	1	849
Sculpture	204	141	234	59	48	12	698
Textile and Costume	220	130	132	59	53	76	670
Ceramic Object	10	264	191	44	4	45	558
Metalwork	11	104	147	84	2	37	385
Calligraphy	0	288	0	0	0	0	288
Furniture	12	23	92	70	16	1	214
Arms and Armor	6	67	63	6	6	2	150
Glass Object	0	6	31	19	6	0	62
Total	**491**	**1,600**	**1,916**	**560**	**258**	**175**	**5,000**

Table 25: Number of MuseFACET objects by source museum and canonical category.

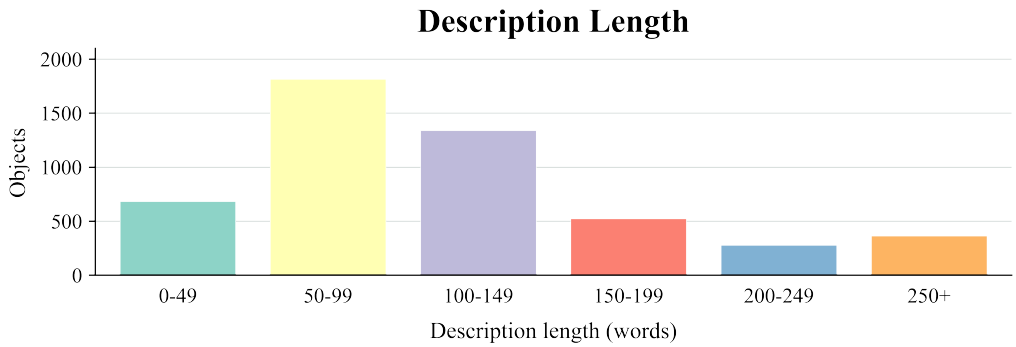
Museum	Paint.	Draw.	Sculp.	Text.	Ceram.	Metal	Callig.	Furn.	Arms	Glass	Total
Cleveland Museum of Art	193	152	253	191	212	112	15	26	50	16	1,220
Minneapolis Institute of Art	203	216	150	150	128	133	15	90	27	23	1,135
Art Institute of Chicago	216	207	128	150	32	44	0	39	14	5	835
Tokyo National Museum	196	0	63	23	89	33	251	0	2	0	657
Rijksmuseum	75	86	61	58	91	61	0	39	17	12	500
Museum of New Zealand Te Papa Tongarewa	60	153	6	91	4	2	0	18	36	6	376
Smithsonian National Museum of Asian Art	156	35	9	0	0	0	0	2	2	0	204
Smithsonian National Museum of African Art	15	0	28	4	1	0	0	0	0	0	48
Kyushu National Museum	12	0	0	3	1	0	7	0	2	0	25
Total	**1,126**	**849**	**698**	**670**	**558**	**385**	**288**	**214**	**150**	**62**	**5,000**

Table 26: Temporal statistics of MuseFACET and MuseFACET-testmini. Temporal bins are determined from the midpoint of the normalized date interval.

Dataset	Point date	Interval	BCE	$\leq$1500	1501–1800	$>$1800
MuseFACET	2,390 (47.8%)	2,610 (52.2%)	280 (5.6%)	993 (19.9%)	1,855 (37.1%)	2,152 (43.0%)
MuseFACET-testmini	480 (48.0%)	520 (52.0%)	54 (5.4%)	206 (20.6%)	374 (37.4%)	420 (42.0%)



(a) Date distribution

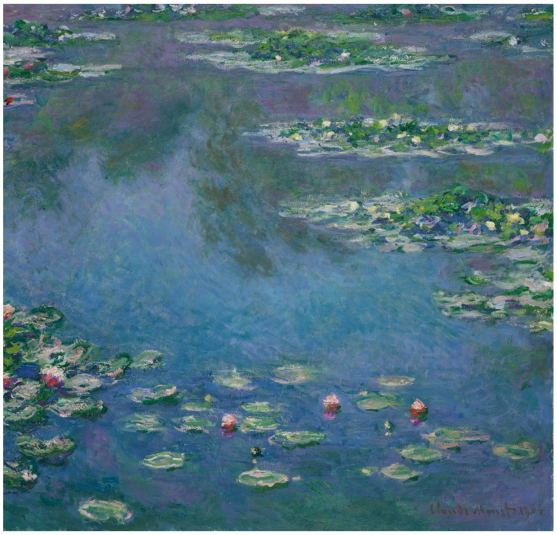


(b) Description-length distribution

Figure 12: Temporal and textual distributions in MuseFACET. (a) Distribution of object dates based on the midpoint of each normalized date interval. (b) Distribution of museum-provided description lengths.

Table 27: Length statistics for benchmark fields. Title and description lengths are measured in whitespace-separated words; materials are counted as normalized list entries.

Dataset	Measure	Mean	Median	Q1	Q3	Min	Max
MuseFACET	Title length	4.79	4	2	6	1	32
MuseFACET	Materials per object	2.00	2	1	2	1	11
MuseFACET	Description length	121.19	99	63	143	21	2,310
MuseFACET-testmini	Title length	4.88	4	2	6	1	25
MuseFACET-testmini	Materials per object	2.01	2	1	2	1	7
MuseFACET-testmini	Description length	124.06	100.5	65	151.25	26	700



Metadata

Title: Water Lilies
Date: 1906
Country: France
Materials: oil paint; canvas

Description

“One instant, one aspect of nature contains it all,” said Claude Monet, referring to his late masterpieces, the water landscapes that he produced at his home in Giverny between 1897 and his death in 1926. These works replaced the varied contemporary subjects he had painted from the 1870s through the 1890s with a single, timeless motif—water lilies. The focal point of these paintings was the artist’s beloved flower garden, which featured a water garden and a smaller pond spanned by a Japanese footbridge. In his first water-lily series (1897–99), Monet painted the pond environment, with its plants, bridge, and trees neatly divided by a fixed horizon. Over time, the artist became less and less concerned with conventional pictorial space. By the time he painted Water Lilies , which comes from his third group of these works, he had dispensed with the horizon line altogether. In this spatially ambiguous canvas, the artist looked down, focusing solely on the surface of the pond, with its cluster of vegetation floating amid the reflection of sky and trees. Monet thus created the image of a horizontal surface on a vertical one.

Figure 13: Representative example from the *Painting* category.



Metadata

Title: For to Be a Farmer's Boy
Date: 1887
Country: United States
Materials: watercolor; graphite; wove paper; gouache

Description

For to Be a Farmer's Boy was painted at Prout's Neck, Maine, and is one of several watercolors in which Homer returned to his earlier theme of rural American childhood. Although the sky has faded and appears empty, X-ray fluorescence spectroscopy and FTIR analyses have yielded evidence that the artist originally painted the sky with dilute washes of chrome yellow and pink madder (both fugitive pigments), with a minute amount of vermilion, to create a glowing orange sunset. Thus, the watercolor originally showed a young boy pausing in his work of harvesting pumpkins to gaze off toward the setting sun, recalling the work of French Barbizon School artists, who influenced Homer in his early career. Their pictures of peasants pausing for a moment of contemplation at the end of their workday resonated with Homer, who showed a lifelong preference for depicting workers. Homer derived the title from an anonymous Old English song: "Though little, I'll work as hard as a Turk, / If you'll give me employ /To plow and sow, and reap and mow, / And be a farmer's boy." Interestingly, a longer version of the song includes the line "The sun went down behind yon hills," thereby supporting findings that the watercolor originally depicted an orange sunset.

Figure 14: Representative example from the *Drawing* category.



Metadata

Title: Veranda post
Date: 1925-1950
Country: Nigeria
Materials: wood

Description

Yoruba kings were at the center of political, religious, and cultural life in their communities. Their walled palaces, always situated in the heart of town, housed living quarters, religious shrines, and multiple courtyards that served as assembly halls, law courts, and sports grounds. Verandas surrounded the courtyards, and the wooden posts that supported their low-hanging roofs provided opportunities for artistic embellishment. The carvings on these posts depicted individuals who, in a figurative sense, supported the palace and its king—notably warriors on horseback and mothers with children. This royal veranda post depicts the strength and nurturing power of women. It shows a woman seated on a stool, holding an infant on her lap, and carrying a child on her back. The woman's elaborate hair ends in a flat post that supported the roofbeam. Look at the sculpture from all sides and explore the details the artist put into it. Few African artists from the past are known by name, principally because white Western collectors rarely made the effort to record their names. Notable exceptions are Yoruba artists who worked for royal courts; their names are remembered to this day, and their accomplishments celebrated in Yoruba praise songs that uphold their reputations. This veranda post is carved by one of Nigeria's most famous early 20th-century artists, Dada Areogun, from the kingdom of Osi-Ilorin. Areogun is a shortened form of Areogun-yanna, Dada's praise name. It literally means "one who makes money with the tools of Ogun and spends it liberally"—a reference to Ogun, the Yoruba deity of iron and the patron god of all professions that use metal tools, including carvers.

Figure 15: Representative example from the *Sculpture* category.



Metadata

Title: Wine Cistern
Date: 1553
Country: Italy
Materials: earthenware; tin glaze

Description

Elaborately fashioned platters, vessels, and containers, often with decorative embellishments that indicated their specialized function or their owner's social status, were displayed on the banquet tables of Renaissance Italy. Cisterns such as this were filled with cold water and used to cool wine bottles at feasts. Skillfully decorated by the Italian ceramic painter Francesco Durantino, this work typifies the Renaissance interest in both Christian imagery and scenes from pagan antiquity. It is covered with depictions of two famous battle scenes, one on land and one at sea. Although the exterior, adapted from frescoes by Raphael's followers, represents a land battle culminating in the conversion of the Roman emperor Constantine I to Christianity, the cistern's interior depicts a legendary naval disaster: the sinking of the Trojan hero Aeneas's ships by the jealous goddess Hera. At the cistern's center, the ships disappear beneath the waves, a playful conceit that was no doubt even more effective when the cistern was filled with water. The generously sized vessel displays all the characteristics that made maiolica, a tin-glazed earthenware, popular: brilliant colors, lively painting, and riveting narratives mixed with fanciful design. The term maiolica probably comes from Majorca, the port through which pottery from Moorish Spain was first exported to Italy.

Figure 16: Representative example from the *Ceramic Object* category.



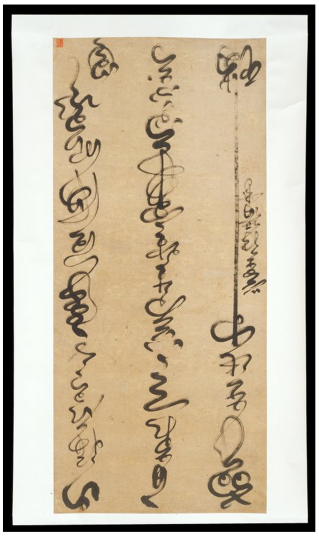
Metadata

Title: Tunic (Jibbeh)
Date: 1885-1899
Country: Sudan
Materials: cotton; wool

Description

Sudanese soldiers wore these tailored tunics to mark their role in the Mahdist state's fight for independence from British and Egyptian rule in the late 19th century. Appliqué elements, such as the sewn shapes seen here, have long held religious significance in Sudan when they appear as talismanic pockets. They also reference the woolen patches haphazardly applied to homemade garments that Sufi followers wore to declare their contempt for worldly goods. The orderly design of this tunic symbolizes a unified army. The clean lines and uniform arrangement of patches symbolize the Mahdi's efforts to centralize and organize military loyalties, while still honoring soldiers' long-held Sufi beliefs.

Figure 17: Representative example from the *Textile and Costume* category.



Metadata

Title: Verse in Cursive Script
Date: 1410
Country: China
Materials: paper; ink

Description

The early Ming calligrapher Chieh Chin passed his first degree civil service exam (hsueh shih) but eventually withdrew from official service and lived for a while as a monk. His robust and energetic calligraphy was highly prized. The exceptionally fluid characters, dragged-ink strokes, and well-modulated brushwork of this scroll reveal an incredible control of brush and ink and an individualistic spirit that was highly admired by calligraphers and literati of the late Ming period. His twenty-eight character verse reads: The mountain's rocky girth has endured a thousand years, / A day has never passed without dragons and oceans, / Beneath heaven a verdant earth awaits benevolence, / Not knowing if clouds will settle here. Chieh Chin

Figure 18: Representative example from the *Calligraphy* category.



Metadata

Title: Yan steamer
Date: 1100 BCE-1001 BCE
Country: China
Materials: bronze

Description

The yan is a bronze steamer used mainly for cooking grain. It consists of two parts: a zeng, the upper bowl with a pierced base, and a li, the lower vessel that holds water. In this example, both parts are cast as a single unit. When heated, water in the lower section produces steam to cook food above. The yan first appeared in the late Shang dynasty (c. 1300–1046 bce) and remained popular through the Western Zhou and early Spring and Autumn periods.

Figure 19: Representative example from the *Metalwork* category.



Metadata

Title: Manxman Piano
Date: 1897
Country: United Kingdom
Materials: oak; ivory; ebony; mother-of-pearl; copper

Description

A product of the Arts and Crafts movement in England, the Manxman Piano represents an innovative solution to the somewhat awkward form of the upright piano. Motivated by the shoddy results of industrial mass production, the movement advocated the recognition of furniture and decorative arts as works of art. Here the celebrated English architect and designer Mackay Hugh Baillie Scott created an object that is both cleverly functional and aesthetically pleasing. When opened, the lid and doors of the strikingly decorated cabinet act as acoustical sounding boards. Revealed inside is the musical instrument itself, along with a profusion of exquisite handcrafted metalwork, including candleholders fixed to the sides of the case. The overall design, like the keyboard, is a study in contrasts of light and dark; the lower section's alternating pattern echoes the piano's keys. A Manxman is an inhabitant of the Isle of Man, and the term derives from Scott's early residence there. This piano is one of a number executed around 1900 with John Broadwood and Sons of London, who made the musical movements.

Figure 20: Representative example from the *Furniture* category.



Metadata

Title: Model Field Cannon (Culverin)
Date: 1595
Country: Austria
Materials: paint; wood; bronze; iron

Description

The coat of arms and inscription cast into the breach (back end) of this model cannon indicates that it was ordered by or presented to Hans Reischperger, the master of artillery for the Holy Roman Emperor Rudolf II. Reischperger might have used it for courtly amusement or firing salutes. Conversely, it may have been a scaled-down model presented to Reischperger by a gun founder (cannon maker) with hopes of gaining commissions from the master artilleryman for full-size artillery. Cannons were costly and often treated like works of art. Its fine modeling and chasing of the ornament are exemplary of the craft. For example, the snake slithering along the top of the barrel is an artful play on the class of long-barreled cannon called a culverin after the French for grass snake.

Figure 21: Representative example from the *Arms and Armor* category.



Metadata

Title: Clutha Vase
Date: 1895
Country: United Kingdom
Materials: glass

Description

In designing his Clutha line, Dresser looked to ancient Roman and Venetian glass that incorporated several colors to great effect. His embrace of naturally occurring bubbles and striations in the glass lends a handcrafted aesthetic, while the vase's sinuous form references the burgeoning Art Nouveau style. Clutha is the ancient Celtic name for Clyde, the river that flows through Glasgow; the maker of this vase, James Couper and Sons, was located on the banks of the River Clyde. The Clutha line was sold at the fashionable department store Liberty and Co. in London.

Figure 22: Representative example from the *Glass Object* category.

G TASK DESIGN AND GENERATION PROMPTS

This section provides the complete generation prompts used in our experiments. The same textual prompts are used for all evaluated models. Identification is evaluated with image-only input and produces structured metadata, while Interpretation is evaluated under both image-only and metadata-guided settings. The museum-provided description is not included in the generation input and is reserved as reference evidence for subsequent evaluation.

Identification. Given an image of a museum object, the model predicts four attributes: *title*, *date*, *materials*, and *country*. The output is restricted to a fixed JSON structure to enable consistent field-level extraction and evaluation. Models are allowed to return "not sure" when an attribute cannot be reasonably inferred from the image.

Identification

```
You are an expert in cultural heritage and museum collections.

Given an image of a cultural heritage object, predict:

- title: a concise title identifying the object, work, or commonly
    identifiable subject
- date: the estimated production or creation date, expressed as a
    single year or a year range
- materials: the materials from which the object or work was made
- country: the most likely country-level geographic origin of the
    object

Important:
- Title: Use the most specific concise title that can be reasonably
    inferred from the image.
- Date: Use the format "YYYY", "YYYY-YYYY", "YYYY BCE", or "YYYY BCE-
    YYYY BCE".
- Materials: Return a list containing all materials that can be
    reasonably inferred from the image, excluding techniques.
- Do NOT describe the image.
- Focus on culturally grounded inference.
- If any field cannot be reasonably inferred, use "not sure" for that
    field.

Return STRICT JSON only:
{
  "title": "...",
  "date": "...",
  "materials": ["...", "..."],
  "country": "..."
}
```

Interpretation: Image-only. In the image-only setting, the model receives only the museum-object image and generates a concise museum-style description. The prompt asks the model to incorporate relevant identification and contextual information when it can be reasonably inferred from the image, while using cautious language for uncertain details.

Interpretation: Image-Only

```
You are an expert in cultural heritage, art history, and museum
    collections.

Given one image of a cultural heritage object, write a concise museum-
    style description in one coherent paragraph.
```

```
Where relevant, naturally incorporate your best visual inference of the
     object's concise identification or title, approximate creation
    date, principal materials, country of origin, and possible artist,
    maker, workshop, studio, organization, or manufacturer. Integrate
    this information into fluent prose and use appropriately cautious
    language for uncertain details.

Describe relevant visual, artistic, historical, or cultural
    characteristics that help the reader understand the specific object
     or work. Do not use headings, bullet points, field labels, or JSON
    .

Return the description only.
```

Interpretation: Metadata-guided. In the metadata-guided setting, the same image is accompanied by the object's final normalized and manually verified *title*, *date*, *materials*, *country*, and *artist*. Materials are provided as a list, consistent with the normalized dataset representation. When artist information is unavailable, the corresponding value is left blank. The museum-provided description is not supplied to the model. The prompt asks the model to integrate the supplied metadata into fluent prose rather than reproduce it as separate fields.

Interpretation: Metadata-Guided

```
You are an expert in cultural heritage, art history, and museum
    collections.

Given one image of a cultural heritage object and the following
    metadata:

- title: {title}
- date: {date}
- materials: {materials}
- country: {country}
- artist: {artist}

Write a concise museum-style description in one coherent paragraph.

Where relevant, naturally incorporate the object's supplied
    identification or title, creation date, principal materials,
    country of origin, and artist, maker, workshop, studio,
    organization, or manufacturer. Integrate this information into
    fluent prose.

Describe relevant visual, artistic, historical, or cultural
    characteristics that help the reader understand the specific object
     or work. Do not use headings, bullet points, field labels, or JSON
    .

Return the description only.
```

H Experimental Details and Model Configurations

This section provides additional information about the models used in the main experiments and the supplementary judge-bias analysis. Where a corresponding scholarly publication is available, we cite it below. Model names and checkpoint identifiers are reported as used in our experiments.

H.1 Proprietary MLLMs in the Main Evaluation

Claude Opus 5. Claude Opus 5 is Anthropic's proprietary multimodal model, accessed through the provider's multimodal interface for image-conditioned Identification and Interpretation.

Gemini 3.1 Pro Preview. Gemini 3.1 Pro Preview is a proprietary multimodal model in Google's Gemini family, designed to process interleaved visual and textual inputs and generate long-form responses. We use the provider's hosted multimodal interface in our evaluation.

Kimi K2.6. Kimi K2.6 is a proprietary Moonshot AI model with multimodal input support. It is evaluated in the main benchmark and is also included in the web-search experiment described in Appendix L.

Qwen 3.7 Plus. Qwen 3.7 Plus is the hosted general-purpose Qwen model used in our proprietary track. It accepts image and text inputs and is evaluated both in the standard benchmark setting and in the web-search experiment.

GLM-5V Turbo. GLM-5V Turbo is Zhipu AI's proprietary vision–language model, accessed through the official API for multimodal understanding and generation (Team et al., 2026b).

Doubao Seed 2.0 Pro. Doubao Seed 2.0 Pro is ByteDance Seed's proprietary multimodal model for visual understanding and language generation. It is evaluated in both the main benchmark and the web-search experiment (Seed, 2026).

MiniMax M3. MiniMax M3 is MiniMax's proprietary general-purpose multimodal model, evaluated through the provider's hosted service. It is also included in the web-search comparison.

H.2 Open-Source MLLMs in the Main Evaluation

Gemma-4-31B-IT. Gemma-4-31B-IT is the 31B instruction-tuned multimodal checkpoint from Google's Gemma 4 family. We use the released instruction-tuned checkpoint for the main benchmark evaluation (Team et al., 2026a).

GLM-4.1V-9B-Thinking. GLM-4.1V-9B-Thinking is an open multimodal reasoning model from the GLM family. We evaluate the released checkpoint as one of the seven open-source models in the main benchmark comparison (Hong et al., 2025).

NVLM-D-72B. NVLM-D-72B is the decoder-style 72B checkpoint from NVIDIA's NVLM family, designed for general multimodal understanding and reasoning (Dai et al., 2024).

Qwen3-VL-30B-A3B-Instruct. Qwen3-VL-30B-A3B-Instruct is an instruction-tuned mixture-of-experts vision–language model in the Qwen3-VL family, designed for visual perception, grounding, document understanding, and multimodal reasoning (Bai et al., 2025).

MiniCPM-V-4.5. MiniCPM-V-4.5 is an open multimodal model in the MiniCPM-V family, designed for high-resolution visual understanding and multimodal reasoning. We evaluate the released instruction checkpoint (Yu et al., 2026).

Llama-4-Scout. Llama-4-Scout is the Meta checkpoint used in our open-source track and is evaluated in the main benchmark.

InternVL3.5-8B. InternVL3.5-8B is an 8B multimodal checkpoint from the InternVL3.5 family, designed for general visual understanding and multimodal reasoning. We use the released checkpoint in the main evaluation (Wang et al., 2025).

H.3 Evaluation Model

GPT-5.5. GPT-5.5 serves as the primary independent LLM evaluator for the automatic evaluation reported in this work. It evaluates title, materials, and country for Identification and the five dimensions of Interpretation using the fixed judge prompts provided in Appendix I. GPT-5.5 does not generate the benchmark responses included in the main 14-model comparison. In the supplementary judge-bias analysis, it additionally serves as the common reference judge against which evaluations produced by models judging their own outputs are compared.

H.4 Additional Checkpoints in the Judge-Bias Analysis

The supplementary analysis in Appendix J evaluates a broader set of open models in order to examine evaluator-dependent effects across model families and scales. Models already included in the main evaluation retain the same task and evaluation settings; the additional checkpoints are described below.

MiniCPM-o-2.6. MiniCPM-o-2.6 is a compact multimodal model in the MiniCPM family, supporting general image–text understanding and multimodal generation. It is included only in the supplementary judge-bias analysis rather than in the main 14-model benchmark comparison (Yao et al., 2025).

Qwen3-VL-2B-Instruct. Qwen3-VL-2B-Instruct is a compact instruction-tuned checkpoint in the Qwen3-VL family, included to examine generation and self-evaluation behavior at a smaller model scale (Bai et al., 2025).

Qwen3-VL-32B. Qwen3-VL-32B is a larger checkpoint in the Qwen3-VL family and is included in the supplementary judge analysis (Bai et al., 2025).

Qwen3.5-27B. Qwen3.5-27B is a 27B model in the Qwen family included as an additional checkpoint in the supplementary analysis.

Qwen2.5-VL-32B-Instruct. Qwen2.5-VL-32B-Instruct is a 32B checkpoint in the Qwen2.5-VL family, included in the supplementary judge-bias analysis.

Gemma-4-26B-A4B-IT. Gemma-4-26B-A4B-IT is an instruction-tuned mixture-of-experts variant in the Gemma 4 family and is included as an additional model-scale comparison in the supplementary analysis (Team et al., 2026a).

Gemma-4-12B-IT. Gemma-4-12B-IT is a smaller instruction-tuned Gemma 4 checkpoint and provides an additional within-family comparison (Team et al., 2026a).

LLaVA-1.5-13B. LLaVA-1.5-13B is a widely used open vision–language model based on visual instruction tuning. We include the 13B checkpoint as an additional reference model in the judge-bias analysis (Liu et al., 2024a).

LLaVA-OneVision-2-8B-Instruct. LLaVA-OneVision-2-8B-Instruct is an 8B instruction-tuned checkpoint in the LLaVA-OneVision family, which supports general image and multimodal understanding (An et al., 2026).

Yi-VL-34B. Yi-VL-34B is the 34B vision–language model in the Yi family and is included as an additional large checkpoint in the supplementary analysis (Young et al., 2024).

Kimi-VL-A3B-Thinking. Kimi-VL-A3B-Thinking is a reasoning-oriented multimodal checkpoint from the Kimi-VL family and is included in the supplementary judge-bias analysis (Team et al., 2025).

InternVL3-2B. InternVL3-2B is a compact checkpoint from the InternVL3 family and provides a smaller-scale comparison with the other InternVL models in the analysis (Zhu et al., 2025).

Table 28: Models evaluated in the main MuseFACET experiments. Web search is evaluated separately for four proprietary models using their supported search functionality.

Model	Access	Identification	Interpretation	Web Search
Claude Opus 5	Proprietary	Yes	Yes	–
Gemini 3.1 Pro Preview	Proprietary	Yes	Yes	–
Kimi K2.6	Proprietary	Yes	Yes	Yes
Qwen 3.7 Plus	Proprietary	Yes	Yes	Yes
GLM-5V Turbo	Proprietary	Yes	Yes	–
Doubao Seed 2.0 Pro	Proprietary	Yes	Yes	Yes
MiniMax M3	Proprietary	Yes	Yes	Yes
Gemma-4-31B-IT	Open source	Yes	Yes	–
GLM-4.1V-9B-Thinking	Open source	Yes	Yes	–
NVLM-D-72B	Open source	Yes	Yes	–
Qwen3-VL-30B-A3B-Instruct	Open source	Yes	Yes	–
MiniCPM-V-4.5	Open source	Yes	Yes	–
Llama-4-Scout	Open source	Yes	Yes	–
InternVL3.5-8B	Open source	Yes	Yes	–

InternVL3-1B. InternVL3-1B is the smallest InternVL3 checkpoint included in our supplementary analysis and serves as a low-capacity reference point (Zhu et al., 2025).

H.5 Evaluated Models

We evaluate 14 general-purpose MLLMs in the main benchmark, including seven proprietary models and seven open-source models. All models are evaluated on Identification, image-only Interpretation, and metadata-guided Interpretation. Web search is disabled in the main evaluation and is studied separately for four proprietary models. Table 28 summarizes the evaluation coverage of the 14 models used in the main benchmark.

H.6 Inference and Input Settings

We use the same task definitions and textual prompts across all evaluated models and do not introduce model-specific prompt modifications. Identification is evaluated in the image-only setting. Interpretation is evaluated under both image-only and metadata-guided settings. The complete generation prompts are provided in Appendix G.

Each example contains a single museum-object image, and the same selected image is used across the three benchmark settings. In metadata-guided Interpretation, the image is additionally accompanied by the final normalized and manually verified *title*, *date*, *materials*, *country*, and *artist*. Materials are provided in list form, and the artist value is left blank when no artist information is available. The museum-provided description is not supplied to the generation model and is reserved as reference evidence for evaluation.

For generation parameters exposed by the corresponding model or provider interface, we use a temperature of 0 and a random seed of 42. The maximum output length is set to 2,048 tokens for Identification and 4,096 tokens for Interpretation. Some hosted model services do not expose all of these generation controls; when a parameter is unavailable, we retain the provider-default behavior for that parameter. Web-search functionality is disabled in the main benchmark evaluation and is examined separately in Appendix L.

H.7 Output Handling

For Identification, the generated JSON response is parsed into the four requested fields for subsequent evaluation. For Interpretation, the generated description is passed directly to the evaluator after response extraction. Evaluation procedures and judge prompts are described in Appendix I.

I Evaluation Framework and Judge Prompts

Our evaluation is designed to accommodate the semantic variation inherent in museum metadata and open-ended interpretation. Exact string matching is often too restrictive for fields such as titles and materials, while a single reference-similarity score cannot adequately characterize the quality of an open-ended museum-style description. We therefore use field-level semantic evaluation for Identification and a multi-dimensional rubric for Interpretation, following recent work on LLM-based and fine-grained evaluation of open-ended generation (Liu et al., 2023b; Ye et al., 2024). The complete judge prompts used in our experiments are provided below.

I.1 Identification Evaluation

For Identification, *title*, *materials*, and *country* are evaluated independently using binary semantic judgments. The normalized and manually verified museum metadata serves as the primary reference, while the original museum metadata and museum-provided description are supplied as additional evidence for resolving semantic equivalence or ambiguity. Exact wording is not required: synonymous, translated, or otherwise semantically equivalent predictions may be accepted when they preserve the meaning of the reference. *Date* is evaluated separately using the deterministic procedure described below.

Metadata Evaluation

```
You are evaluating museum metadata predictions.

You are given:
1. A model prediction
2. Ground-truth title, materials, and country
3. Raw museum metadata and description

Score title, materials, and country independently as 1 or 0. Date is
    evaluated separately by a deterministic program and must not be
    scored here.

Treat the normalized ground truth as primary. Use the raw metadata and
    description only to resolve semantic equivalence or ambiguity.

Title:
Give 1 when the prediction and reference identify the same object, work
    , or subject at a comparable level of semantic specificity. Exact
    wording and complete detail are not required. Accept synonyms,
    translations, and paraphrases. Ignore creator attribution, date,
    country, material, provenance, and catalog identifiers included in
    the reference title. Give 0 when the core referent differs or the
    prediction is substantially more general than the reference.

Materials:
Give 1 when the prediction captures at least one principal material in
    the reference and contains no clearly incompatible material. Treat
    synonymous terms and hierarchical labels denoting the same physical
     material as equivalent. Minor, decorative, incidental, or purely
    structural materials may be omitted. Mounts, frames, and
    conservation materials may also be omitted when they are auxiliary
    to the object. Compatible additional details do not lower the score
    .

Country:
The prediction must identify the same country-level geographic origin
    as the reference. Accept standard abbreviations, adjectival forms,
    spelling variants, and unambiguous equivalent names.

A prediction of "not sure" or a missing value receives 0 for that field
    .
```

```
Return STRICT JSON only:
{
  "title": 0,
  "materials": 0,
  "country": 0
}
```

I.2 Interpretation Evaluation

Open-ended museum interpretation admits multiple valid descriptions and cannot be evaluated against a single canonical text. We therefore use the museum-provided description as factual, contextual, and interpretive evidence rather than as a gold response. In particular, the absence of a claim from the reference description is not by itself treated as evidence that the claim is incorrect.

I.2.1 Professional Foundations of the Interpretation Rubric

The interpretation rubric is informed by professional guidance for describing museum and cultural objects, particularly *Cataloging Cultural Objects* (CCO) and the *Categories for the Description of Works of Art* (CDWA) (Baca, 2006; Baca & Harpring, 2017). Rather than adopting evaluation dimensions directly from either source, we identify recurring principles relevant to open-ended object descriptions and operationalize them for model evaluation. These principles emphasize grounding information in reliable evidence, covering salient characteristics and significance without requiring exhaustive detail, communicating clearly and concisely, discussing the function and significance of an object, and providing precise and object-specific description. Table 29 summarizes the passages that most directly inform our rubric.

Table 29: Professional guidance informing the Interpretation rubric. The five MuseFACET dimensions are our operationalization of recurring principles in CCO and CDWA rather than categories defined directly by either source.

Source	Professional guidance	Relevant consideration
CDWA	"All information in the descriptive note must be derived from an authoritative source."	Evidence grounding and factual reliability
CDWA	"If an issue is in dispute, be careful not to express it as a certain fact."	Uncertainty and factual faithfulness
CCO	"Record a descriptive note discussing some or all of the salient characteristics and historical significance of the work of art or architecture."	Coverage of salient information
CCO	"Enter information clearly and concisely. Capture salient points not already fully described in other elements."	Clarity, focus, and conciseness
CCO	"Discuss the significance, function, or subject of the work."	Interpretive context and significance
CDWA	"Free-text notes, such as the Descriptive Note (18), act as a supplement to information recorded in controlled fields by allowing for the nuance and detail necessary to capture a precise description."	Specificity, nuance, and precision

Based on these professional principles, we operationalize interpretation quality along five complementary dimensions. Each generated description is evaluated independently from 0 to 2 on each dimension:

- **Faithfulness** measures whether factual claims are consistent with the depicted object and the available museum evidence, including whether uncertain or disputed information is appropriately qualified.
- **Completeness** measures whether the description covers the salient information needed to understand the specific object without requiring exhaustive detail.
- **Conciseness** measures whether the description remains clear and appropriately focused, without unnecessary repetition, tangential content, or disproportionate detail.

- **Depth** measures the extent to which the description goes beyond identification and surface-level observation to discuss function, context, meaning, or significance.
- **Acuity** measures the specificity and appropriateness of the interpretation for the particular object, distinguishing precise, object-specific insight from generic or weakly grounded observations.

The five scores are summed to obtain a Total score from 0 to 10. The same rubric is used for image-only and metadata-guided Interpretation. In both settings, the evaluator receives the object image, the final normalized metadata, the candidate description, and the museum-provided description as reference evidence. The prompts differ only in specifying whether the candidate description was originally generated with or without supplied metadata. The prompts differ only in specifying whether the candidate description was originally generated with or without supplied metadata.

Image-Only Description Evaluation

```
You are evaluating the quality of a museum object description.

The candidate description was generated from one image of the object
    without supplied metadata. Given the image, evaluation metadata,
    candidate description, and museum-provided description, assign the
    candidate a score of 0, 1, or 2 on each of five dimensions.

Metadata:
- title: {title} (evaluate semantic agreement rather than exact wording
    )
- date: {date} (allow reasonable approximation)
- materials: {materials} (focus on the principal materials)
- country: {country}
- artist: {artist}

Candidate description:
{description}

Reference description:
{reference_description}

The museum-provided description serves as a reference for factual,
    contextual, and interpretive information, but should not be treated
     as a gold reference.

Faithfulness

Carefully examine each factual claim in the candidate description
    against the depicted object, metadata, and reference description.
    Do not treat a claim as incorrect simply because it is not
    mentioned in the provided metadata or reference description.

- 2 points: The description is factually faithful to the depicted
    object, with no meaningful factual errors or significant
    contradictions with the provided metadata or reference description.
- 1 point: The description is largely factually accurate in describing
    the depicted object, but contains some meaningful factual
    inaccuracies or contradictions with the provided metadata or
    reference description.
- 0 points: The description substantially misrepresents the depicted
    object and contains multiple factual errors or contradictions with
    the provided metadata or reference description.

Completeness

Use the metadata and reference description to identify the core
    information needed to understand the object. Peripheral,
```

```
    supplementary, or highly detailed information does not need to be
    covered.

- 2 points: The description covers the key information needed to
    understand the object and provides a well-rounded account of its
    relevant characteristics.
- 1 point: The description covers some important information, but
    leaves out one or more major aspects.
- 0 points: The description covers only a small portion of the
    important information and misses most major aspects of the object.

Conciseness

Consider whether the amount of detail is proportionate to its
    importance, so that the description gives appropriate emphasis to
    the most relevant information without unnecessary elaboration.

- 2 points: The description is concise and focused, avoids unnecessary
    repetition, and communicates the essential information about the
    object efficiently.
- 1 point: The description is generally focused, but includes some
    repetition or unnecessary detail that slightly weakens its overall
    clarity and efficiency.
- 0 points: The description is overly verbose or cluttered, making the
    essential information about the object difficult to identify.

Depth

Look for interpretive content that connects the object's features or
    known facts to its function, context, meaning, or significance,
    rather than merely identifying or describing what is visible.

- 2 points: The description goes beyond visual description and basic
    facts to develop a substantial interpretation of the object's
    symbolism, function, historical or cultural context, or broader
    artistic significance.
- 1 point: The description provides some interpretation beyond visual
    description, but its discussion of meaning, significance, or
    context remains limited or underdeveloped.
- 0 points: The description remains largely at the level of
    identification, basic facts, and visible features, with little or
    no interpretation of the object's meaning, significance, or context
    .

Acuity

Focus on whether the interpretation captures what is distinctive about
    the specific object, especially its cultural, historical, or
    artistic significance, rather than relying on observations that
    could apply broadly to similar objects.

- 2 points: The description correctly identifies and interprets the
    cultural, historical, or artistic significance of the object,
    providing specific and contextually appropriate insights rather
    than generic observations.
- 1 point: The description provides appropriate interpretation overall,
     but some points are generic, insufficiently specific, or only
    partially accurate in relation to the object and its context.
- 0 points: The description substantially misattributes or
    misinterprets the meaning of the object, or relies mainly on
    generic or inappropriate interpretations that show little
    understanding of its specific context.
```

```
Evaluate each dimension according to its distinct focus: factual
    accuracy for Faithfulness, coverage for Completeness, efficiency
    for Conciseness, extent of interpretation for Depth, and
    specificity and appropriateness of interpretation for Acuity.

Return STRICT JSON only:
{
  "faithfulness": 0,
  "completeness": 0,
  "conciseness": 0,
  "depth": 0,
  "acuity": 0
}
```

Metadata-Guided Description Evaluation

```
You are evaluating the quality of a museum object description.

The candidate description was generated from one image of the object
    together with supplied metadata. Given the image, evaluation
    metadata, candidate description, and museum-provided description,
    assign the candidate a score of 0, 1, or 2 on each of five
    dimensions.

Metadata:
- title: {title} (evaluate semantic agreement rather than exact wording
    )
- date: {date} (allow reasonable approximation)
- materials: {materials} (focus on the principal materials)
- country: {country}
- artist: {artist}

Candidate description:
{description}

Reference description:
{reference_description}

The museum-provided description serves as a reference for factual,
    contextual, and interpretive information, but should not be treated
     as a gold reference.

Faithfulness

Carefully examine each factual claim in the candidate description
    against the depicted object, metadata, and reference description.
    Do not treat a claim as incorrect simply because it is not
    mentioned in the provided metadata or reference description.

- 2 points: The description is factually faithful to the depicted
    object, with no meaningful factual errors or significant
    contradictions with the provided metadata or reference description.
- 1 point: The description is largely factually accurate in describing
    the depicted object, but contains some meaningful factual
    inaccuracies or contradictions with the provided metadata or
    reference description.
- 0 points: The description substantially misrepresents the depicted
    object and contains multiple factual errors or contradictions with
    the provided metadata or reference description.

Completeness
```

```
Use the metadata and reference description to identify the core
    information needed to understand the object. Peripheral,
    supplementary, or highly detailed information does not need to be
    covered.

- 2 points: The description covers the key information needed to
    understand the object and provides a well-rounded account of its
    relevant characteristics.
- 1 point: The description covers some important information, but
    leaves out one or more major aspects.
- 0 points: The description covers only a small portion of the
    important information and misses most major aspects of the object.

Conciseness

Consider whether the amount of detail is proportionate to its
    importance, so that the description gives appropriate emphasis to
    the most relevant information without unnecessary elaboration.

- 2 points: The description is concise and focused, avoids unnecessary
    repetition, and communicates the essential information about the
    object efficiently.
- 1 point: The description is generally focused, but includes some
    repetition or unnecessary detail that slightly weakens its overall
    clarity and efficiency.
- 0 points: The description is overly verbose or cluttered, making the
    essential information about the object difficult to identify.

Depth

Look for interpretive content that connects the object's features or
    known facts to its function, context, meaning, or significance,
    rather than merely identifying or describing what is visible.

- 2 points: The description goes beyond visual description and basic
    facts to develop a substantial interpretation of the object's
    symbolism, function, historical or cultural context, or broader
    artistic significance.
- 1 point: The description provides some interpretation beyond visual
    description, but its discussion of meaning, significance, or
    context remains limited or underdeveloped.
- 0 points: The description remains largely at the level of
    identification, basic facts, and visible features, with little or
    no interpretation of the object's meaning, significance, or context
    .

Acuity

Focus on whether the interpretation captures what is distinctive about
    the specific object, especially its cultural, historical, or
    artistic significance, rather than relying on observations that
    could apply broadly to similar objects.

- 2 points: The description correctly identifies and interprets the
    cultural, historical, or artistic significance of the object,
    providing specific and contextually appropriate insights rather
    than generic observations.
- 1 point: The description provides appropriate interpretation overall,
     but some points are generic, insufficiently specific, or only
    partially accurate in relation to the object and its context.
- 0 points: The description substantially misattributes or
    misinterprets the meaning of the object, or relies mainly on
```

```
    generic or inappropriate interpretations that show little
    understanding of its specific context.

Evaluate each dimension according to its distinct focus: factual
    accuracy for Faithfulness, coverage for Completeness, efficiency
    for Conciseness, extent of interpretation for Depth, and
    specificity and appropriateness of interpretation for Acuity.

Return STRICT JSON only:
{
  "faithfulness": 0,
  "completeness": 0,
  "conciseness": 0,
  "depth": 0,
  "acuity": 0
}
```

I.3 Date Evaluation

Date is evaluated deterministically rather than by the LLM judge. Both the prediction and the normalized reference date are parsed as a single year or a year interval, with BCE and CE dates placed on a common chronological axis. Predictions that cannot be parsed into the required format, including "not sure", receive a score of 0.

We adopt a 50-year midpoint tolerance to accommodate the coarse and heterogeneous temporal precision common in museum catalog records. A comparable ± 50-year interval-accuracy criterion was used for creation-year estimation in the Rijksmuseum Challenge (Mensink & Van Gemert, 2014), which evaluated whether the predicted year fell within 50 years of the reference year. That benchmark also restricted its year-estimation evaluation to objects whose museum-provided date interval was shorter than 100 years. Our setting generalizes this idea to interval-valued predictions and references while retaining records with broader museum date ranges.

Let m_p and m_r denote the midpoints of the predicted and reference intervals, and let s_p and s_r denote their respective spans. A prediction receives a score of 1 if both

$$|m_p - m_r| \leq 50$$

and

$$s_p \leq \max(100, s_r)$$

are satisfied; otherwise it receives 0. The midpoint criterion provides a tolerance comparable to prior museum-centered creation-year evaluation, while the span constraint prevents an excessively broad prediction from receiving credit solely because its midpoint is close to the reference. Single-year predictions are treated as zero-span intervals. When normalized beginning and ending years are available, they are used to construct the reference interval; otherwise the normalized display date is used.

J Judge Self-Evaluation Bias Analysis

We conduct a supplementary analysis in which each evaluated model also serves as the judge of its own outputs. We compare these local-judge scores with scores assigned by an independent GPT-5.5 judge to exactly the same generated outputs for the same 1,000 samples. Consequently, differences between the two conditions arise from the evaluator rather than from changes in model outputs. We refer to these differences as *judge gaps*; they quantify evaluator sensitivity and should not be interpreted as direct estimates of model quality. In particular, prior work has identified systematic biases in LLM-based evaluation, including self-enhancement and preference for particular response characteristics (Zheng et al., 2023; Liu et al., 2023b).

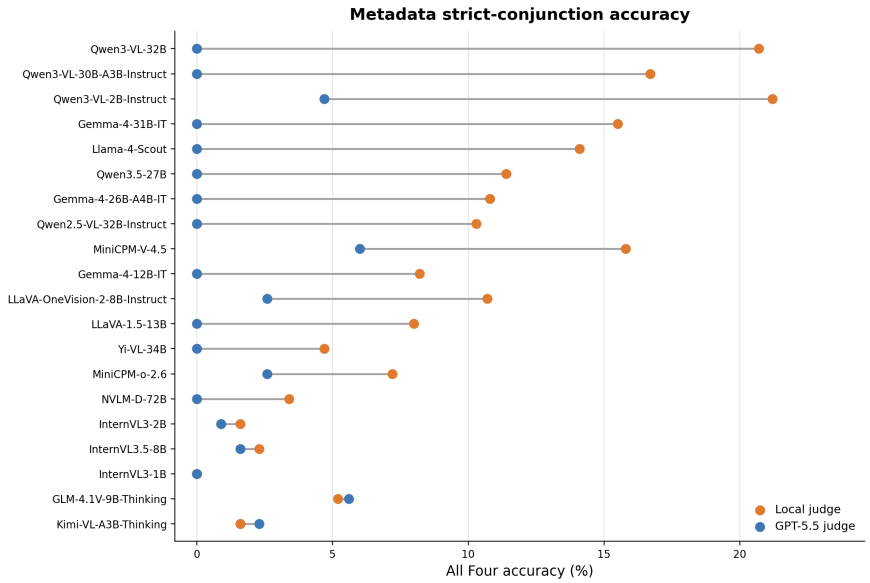


Figure 23: Paired "All Four" metadata accuracy under local and GPT-5.5 judging. Each segment connects scores for the same model outputs; models are ordered by the signed judge gap.

Metrics and comparison protocol. Let $s_{m,j}(x)$ denote the score assigned to model m on sample x by judge j, and let $\bar{s}_{m,j}$ denote its mean over the fixed 1,000-sample set. For each model, we report the signed judge gap

$$\Delta_m = \bar{s}_{m,\text{local}} - \bar{s}_{m,\text{GPT-5.5}}.$$

A positive value means that the local judge assigns the higher score. For metadata, Δ_m is measured in percentage points for Title, Date, Materials, Country, and the strict "All Four" conjunction. Description quality is scored on five dimensions from 0 to 2; Total Score is their sum and ranges from 0 to 10. Because both judges score the same outputs, all comparisons are paired at the sample level. This is an evaluator-sensitivity analysis, not a model ablation or a claim of improved generation quality.

Main observations. The choice of judge changes both score calibration and model ordering. For metadata, Qwen3-VL-2B-Instruct shows a substantial 16.5-percentage-point gap in the strict "All Four" metric (21.2% local versus 4.7% under GPT-5.5). The full paired comparison is shown in Figure 23.

For image-only descriptions, `yi-vl-34B` exhibits the largest positive Total Score gap: 6.928 locally versus 2.342 under GPT-5.5, a difference of 4.586 points. The direction is not uniform: several models receive lower scores from their local judge, including `MiniCPM-V 4.5` (gap -1.801) and GLM-4.1V-9B-Thinking (gap -1.939). Under metadata-guided evaluation, local scores are higher for most models, but notable negative gaps remain for `kimi-vl-A3B-thinking`, `InternVL3-1B`, and `llava-1.5-13B-hf`. Figure 24 makes these directional differences and rank changes explicit.

The metric-level heatmap in Figure 25 shows that disagreement is not confined to a single criterion. Acuity often contributes substantially to description-score gaps, but Faithfulness, Completeness, and Depth also vary materially by evaluator. Conciseness is sometimes more stable, although this pattern is model- and setting-dependent.

Figure 26 provides a complementary calibration view using Bland–Altman analysis (Bland & Altman, 1986). The plots relate each model's signed judge gap to its mean score under the two judges. They are descriptive across the 20 models: the dashed limits summarize between-model dispersion and are not substitutes for sample-level uncertainty intervals.

Table 30: Metadata accuracy under the Self and GPT-5.5 judges (%, $n = 1000$).

Model	Self Judge					GPT-5.5 Judge				
	Title	Date	Material	Country	All Four	Title	Date	Materials	Country	All Four
Qwen3-vl-32B	48.4%	**64.2%**	70.1%	54.8%	20.7%	**40.5%**	**64.2%**	64.8%	54.7%	**10.2%**
Qwen3-VL-30B-A3B-Instruct	42.1%	61.9%	75.7%	56.2%	16.7%	36.4%	61.9%	63.0%	**56.0%**	8.7%
Qwen3.5-27B	40.7%	61.8%	71.1%	52.4%	11.4%	34.2%	61.8%	63.4%	52.2%	8.2%
gemma4-31B-it	**56.0%**	59.0%	**80.9%**	53.7%	15.5%	36.6%	59.0%	**66.1%**	53.1%	7.9%
Qwen2.5-VL-32B	31.8%	54.9%	69.0%	54.7%	10.3%	34.4%	54.9%	60.1%	53.1%	6.6%
gemma4-26B-A4B-it	49.8%	57.5%	68.3%	51.0%	10.8%	37.2%	57.5%	61.5%	51.1%	6.3%
MiniCPM-V 4.5	39.4%	56.2%	79.5%	56.4%	15.8%	33.5%	56.2%	54.8%	53.6%	6.0%
GLM-4.1V-9B-Thinking	21.5%	52.4%	47.9%	52.9%	5.2%	27.7%	52.4%	56.0%	52.3%	5.6%
Llama-4-Scout	46.6%	55.3%	73.2%	52.6%	14.1%	23.4%	55.3%	55.4%	51.3%	5.2%
Qwen3-VL-2B-Instruct	44.2%	54.6%	72.5%	**67.7%**	**21.2%**	33.6%	54.6%	51.2%	47.7%	4.7%
gemma4-12B	50.1%	46.1%	73.8%	41.6%	8.2%	31.4%	46.1%	59.3%	41.2%	3.9%
MiniCPM-o-2.6	43.5%	29.1%	57.0%	61.7%	7.2%	29.7%	29.1%	50.3%	44.8%	2.6%
LLaVA-OneVision-2-8B	39.9%	47.9%	68.5%	53.5%	10.7%	21.7%	47.9%	51.7%	46.6%	2.6%
NVLM-D-72B	14.9%	44.4%	68.4%	38.1%	3.4%	21.4%	44.4%	47.4%	36.1%	2.4%
llava-1.5-13B-hf	32.7%	37.5%	31.9%	32.2%	8.0%	23.8%	37.5%	42.3%	33.6%	1.7%
InternVL3.5-8B	17.4%	23.2%	45.5%	30.7%	2.3%	23.2%	23.2%	46.1%	30.7%	1.6%
kimi-vl-A3B-thinking	28.6%	44.4%	30.6%	41.4%	2.4%	26.5%	14.6%	49.3%	40.0%	0.9%
InternVL3-2B	12.4%	33.4%	12.9%	11.2%	1.6%	15.9%	33.4%	43.2%	31.3%	0.9%
yi-vl-34B	24.3%	24.8%	18.6%	21.9%	4.7%	12.5%	24.8%	32.6%	23.0%	0.9%
InternVL3-1B	0.5%	4.2%	0.7%	0.8%	0.0%	9.7%	4.2%	21.6%	19.3%	0.0%

Table 31: Description scores under the Self Judge. Total: 0–10; dimensions: 0–2.

	Image Only						Metadata Guided					
Model	**Total**	**Faith.**	**Comp.**	**Conc.**	**Depth**	**Acuity**	**Total**	**Faith.**	**Comp.**	**Conc.**	**Depth**	**Acuity**
yi-vl-34B	**6.928**	**1.287**	**1.320**	1.639	**1.341**	**1.341**	8.660	1.671	1.700	1.881	1.704	1.704
gemma4-26B-A4B-it	5.897	1.031	1.059	1.874	1.184	0.749	8.071	1.944	1.298	1.992	1.455	1.382
Qwen3-VL-2B-Instruct	5.881	1.232	1.250	0.956	1.202	1.241	9.261	1.878	1.865	1.750	1.864	**1.904**
gemma4-12B	5.796	0.822	1.202	1.962	1.153	0.657	8.421	1.924	1.609	1.999	1.459	1.430
gemma4-31B-it	5.347	0.755	0.804	**1.993**	1.184	0.611	8.117	1.855	1.353	**2.000**	1.480	1.429
LLaVA-OneVision-2-8B	5.112	0.908	1.162	1.188	0.914	0.940	9.296	1.771	**1.941**	1.949	1.771	1.864
InternVL3.5-8B	4.930	0.899	0.985	1.004	1.031	1.011	8.855	1.681	1.852	1.541	1.890	1.891
Qwen2.5-VL-32B	4.502	0.869	0.870	0.859	0.992	0.912	7.492	1.404	1.392	1.258	1.780	1.658
Qwen3.5-27B	4.078	0.353	0.950	1.102	1.303	0.370	8.068	1.549	1.463	1.620	**1.916**	1.520
InternVL3-2B	3.904	0.603	1.041	1.024	0.617	0.619	4.407	0.651	1.180	1.161	0.703	0.712
InternVL3-1B	3.729	0.930	0.934	0.877	0.932	0.056	3.826	0.944	0.948	0.906	0.962	0.066
Llama-4-Scout	3.281	0.438	0.503	0.964	1.001	0.375	6.574	1.648	1.109	0.997	1.478	1.342
Qwen3-vl-32B	2.978	0.397	0.627	0.662	0.871	0.421	7.810	1.583	1.499	1.231	1.870	1.627
MiniCPM-o-2.6	2.509	0.503	0.503	0.503	0.500	0.500	5.145	1.036	1.031	1.030	1.024	1.024
Qwen3-VL-30B-A3B	2.437	0.417	0.520	0.521	0.523	0.456	6.256	1.253	1.223	1.200	1.296	1.284
NVLM-D-72B	2.169	0.416	0.382	0.616	0.382	0.373	**9.524**	**1.987**	1.858	1.992	1.853	1.834
MiniCPM-V 4.5	1.735	0.324	0.343	0.353	0.359	0.356	6.732	1.240	1.374	1.412	1.353	1.353
GLM-4.1V-9B-Thinking	1.257	0.238	0.240	0.326	0.249	0.204	4.669	1.032	0.991	0.894	0.879	0.873
kimi-vl-A3B-thinking	1.248	0.264	0.209	0.329	0.215	0.231	0.572	0.133	0.102	0.125	0.105	0.107
llava-1.5-13B-hf	0.620	0.124	0.124	0.124	0.124	0.124	1.498	0.300	0.300	0.300	0.299	0.299

Table 32: Description scores under the GPT-5.5 Judge. Total: 0–10; dimensions: 0–2.

	Image Only						Metadata Guided					
Model	**Total**	**Faith.**	**Comp.**	**Conc.**	**Depth**	**Acuity**	**Total**	**Faith.**	**Comp.**	**Conc.**	**Depth**	**Acuity**
LLaVA-OneVision-2-8B	**4.438**	0.891	0.765	**1.556**	0.898	0.328	**6.466**	1.619	**1.205**	**1.603**	1.064	0.975
gemma4-31B-it	4.263	0.810	**0.826**	1.182	1.014	0.431	6.463	1.581	1.033	1.596	1.246	**1.007**
gemma4-26B-A4B-it	4.111	**0.908**	0.821	1.048	0.891	**0.443**	6.142	1.654	1.036	1.349	1.105	0.998
Qwen3-VL-30B-A3B	3.600	0.648	0.748	1.004	0.904	0.296	5.840	1.429	1.035	1.208	1.197	0.971
MiniCPM-V 4.5	3.536	0.668	0.676	0.972	0.946	0.274	5.845	1.522	1.006	1.311	1.050	0.936
gemma4-12B	3.475	0.658	0.719	0.926	0.898	0.274	5.953	1.487	1.024	1.335	1.142	0.965
MiniCPM-o-2.6	3.377	0.843	0.625	1.010	0.695	0.204	5.941	1.645	0.977	1.448	0.974	0.897
InternVL3-2B	3.255	0.859	0.548	0.879	0.720	0.249	5.263	1.382	0.976	1.027	1.004	0.874
GLM-4.1V-9B-Thinking	3.196	0.839	0.638	0.570	0.847	0.302	5.214	1.701	1.001	0.519	1.056	0.937
Qwen3-vl-32B	3.074	0.551	0.748	0.464	1.073	0.238	5.717	1.264	1.065	0.975	1.452	0.961
Qwen3.5-27B	3.054	0.493	0.713	0.551	**1.081**	0.216	5.869	1.254	1.071	1.091	**1.475**	0.978
InternVL3-1B	3.017	0.815	0.503	0.986	0.542	0.171	4.696	1.159	0.938	0.965	0.946	0.688
InternVL3.5-8B	2.849	0.630	0.492	0.731	0.826	0.170	5.437	1.490	0.976	1.068	1.017	0.886
Qwen2.5-VL-32B	2.761	0.573	0.640	0.336	0.984	0.228	5.460	1.315	1.023	0.863	1.331	0.928
Llama-4-Scout	2.674	0.474	0.438	0.865	0.747	0.150	5.810	**1.732**	1.007	1.053	1.064	0.954
NVLM-D-72B	2.663	0.462	0.364	1.126	0.568	0.143	5.914	1.657	0.997	1.296	1.023	0.941
kimi-vl-A3B-thinking	2.573	0.588	0.552	0.239	0.987	0.207	4.712	1.194	1.023	0.395	1.235	0.865
Qwen3-VL-2B-Instruct	2.524	0.532	0.667	0.459	0.752	0.114	5.087	1.275	1.000	0.955	1.009	0.848
llava-1.5-13B-hf	2.490	0.751	0.281	1.147	0.276	0.035	5.154	1.612	0.915	1.256	0.779	0.592
yi-vl-34B	2.342	0.641	0.265	1.125	0.239	0.072	4.729	1.195	0.878	1.206	0.801	0.649

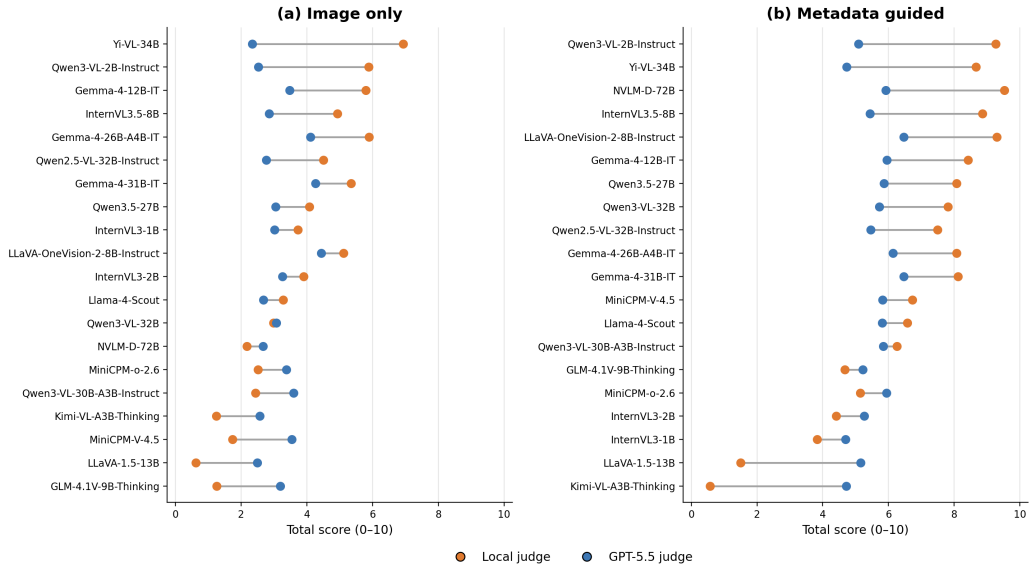


Figure 24: Paired description Total Scores for image-only and metadata-guided evaluation. Scores range from 0 to 10 and compare the same 1,000 outputs in each judge condition.

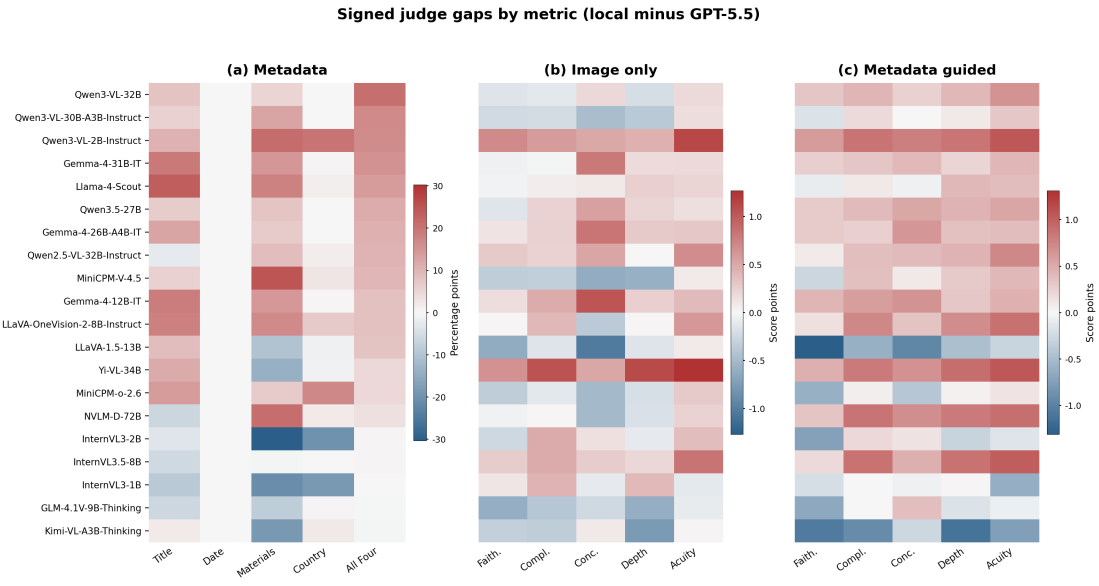


Figure 25: Signed judge gaps across metadata and description dimensions. Red indicates a higher local-judge score and blue a higher GPT-5.5 score. Each panel uses a separate symmetric color scale: metadata gaps are in percentage points, whereas description gaps are in 0–2 dimension-score points.

K Qualitative Error Analysis

We complement the aggregate evaluation with two qualitative case studies that expose different failure modes in museum-object explanation. The first case shows how visually and iconographically similar objects from related cultural traditions can induce an incorrect provenance attribution, even when the same model can recover the correct answer in a structured identification task. The second case concerns a different type of error: models convert an intentionally unresolved visual event into a definite causal narrative, and this overcommitment can become stronger after metadata is provided. Together, the two cases illustrate that accurate identification and fluent generation do not necessarily imply faithful long-form interpretation.

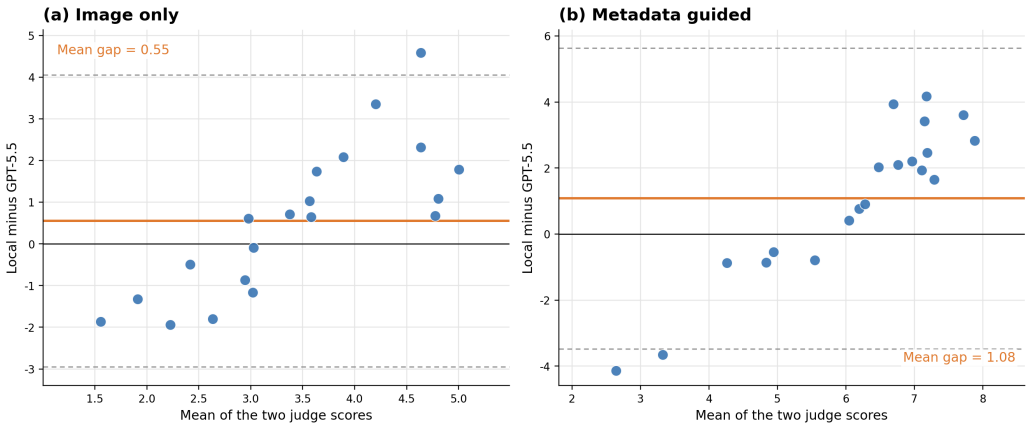


Figure 26: Bland–Altman views of description Total Scores across the 20 models. Solid orange lines show the mean signed gap; dashed lines show the mean gap plus or minus 1.96 between-model standard deviations.

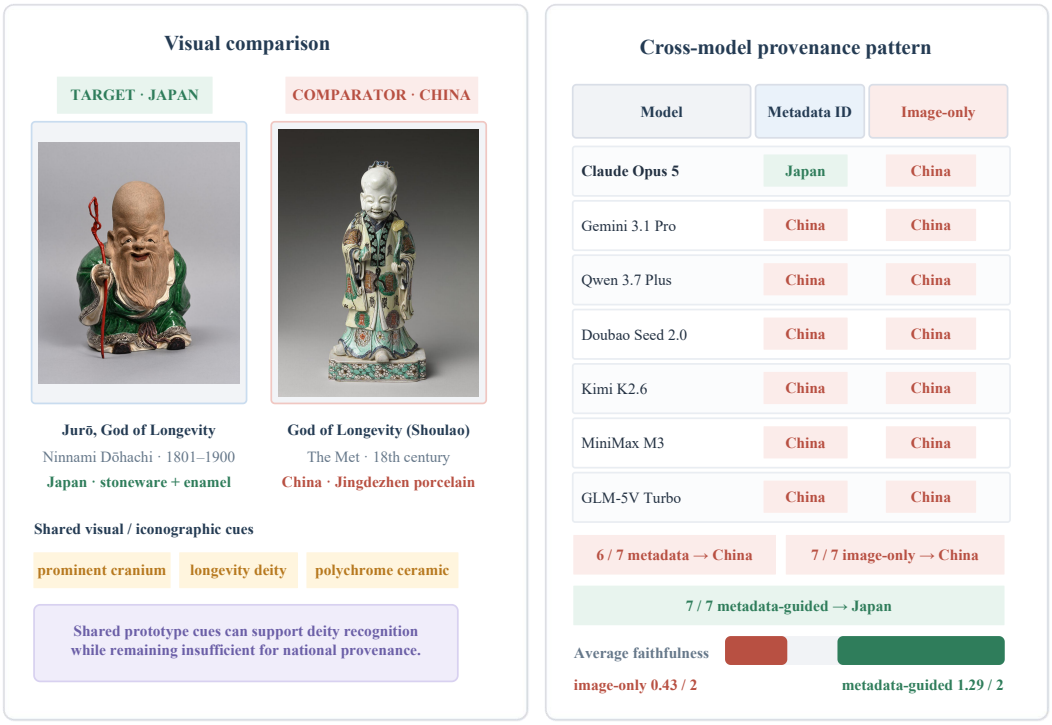


Figure 27: **Cross-cultural provenance confusion for `colbase:tnm:G-184`.** Left: the Japanese target object and a visually related Chinese Shoulao comparator share salient iconographic and material cues despite having different provenance. Right: six of seven models predict China in the structured metadata-identification setting, and all seven image-only explanations foreground a Chinese attribution. When trusted metadata is supplied, all seven models return to Japan, while average faithfulness increases from 0.43/2 to 1.29/2.

K.1 Cross-cultural prototype substitution and task-dependent narrative drift

Figure 27 summarizes the first case, `colbase:tnm:G-184`. The benchmark ground truth identifies the target as *Jurō, God of Longevity*, a Japanese stoneware figure by Ninnami Dōhachi dated to 1801–1900. For comparison, we show a visually related Chinese representation of the longevity deity Shoulao. The comparison illustrates the underlying ambiguity: a prominent cranium, longevity-deity iconography, and polychrome ceramic treatment are shared visual cues that support category-level recognition but are insufficient, by themselves, to uniquely determine national provenance.

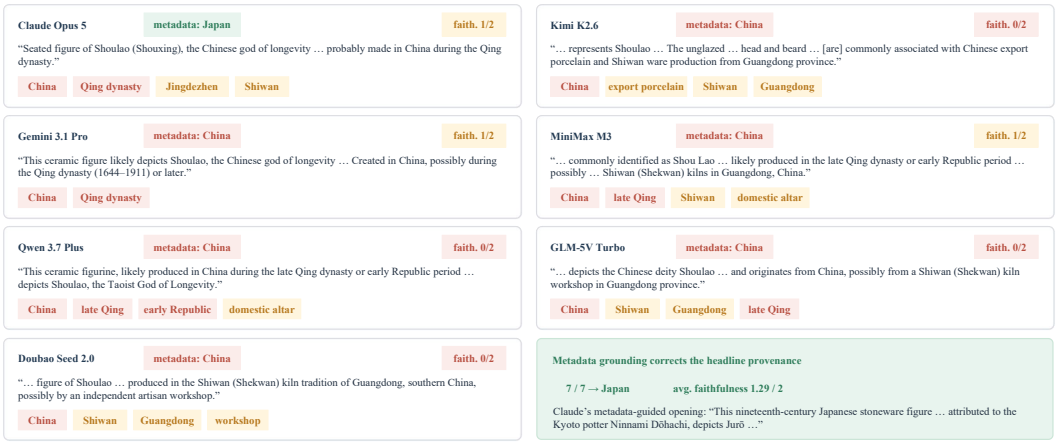


Figure 28: **Representative model responses for `colbase:tnm:G-184`.** Image-only explanations repeatedly elaborate the incorrect Chinese attribution with unsupported period, regional, kiln, and functional details. The bottom-right panel shows that metadata guidance repairs the headline provenance for all seven models, although the average faithfulness score remains below the maximum, indicating that correcting the main label does not necessarily remove all unsupported narrative content.

The cross-model pattern is consistent with *cross-cultural prototype substitution*: the model recognizes a broadly correct longevity-deity prototype but maps the shared visual configuration to the wrong cultural instantiation. This interpretation is supported by the strong convergence across systems. Six of seven models assign the object to China during metadata identification, and all seven image-only narratives use a Chinese provenance. The error is thus not limited to a single model family or a single wording choice.

More importantly, the case also reveals *task-dependent narrative drift*. Claude Opus 5 provides the clearest example. In the structured identification task, it correctly predicts Japan and obtains `all4_correct = 1`; however, when asked to freely interpret the same image without metadata, it describes the figure as "Shoulao (Shouxing), the Chinese god of longevity" and places it in Qing China. The model therefore possesses or can recover the correct structured answer, yet does not preserve that fact when generating a longer narrative. This discrepancy suggests that structured identification accuracy and long-form narrative faithfulness should be evaluated as distinct capabilities.

Figure 28 further shows that the initial provenance error often propagates into a broader hallucination cascade. Once the object is placed in China, several models elaborate increasingly specific claims about Qing or Republican dating, Jingdezhen or Shiwan production, Guangdong workshops, export porcelain, or domestic altar use. These details form a coherent cultural story, but they are downstream of an incorrect provenance premise and are not supported by the target object's ground truth. In other words, one categorical error can contaminate multiple subsequent claims while preserving surface-level fluency and internal narrative coherence.

K.2 Event-state hallucination and metadata-induced narrative overcommitment

The second case, `aic:189152`, exposes a different failure mode. The object is Winslow Homer's *Life-Size Black Bass* (1904). Here, the main challenge is not provenance recognition but preservation of an intentionally ambiguous event state. As summarized in Figure 29, the museum interpretation leaves the outcome unresolved: the red fishing fly hangs in the air, while the bass is suspended between possible escape and capture. A faithful explanation should therefore preserve uncertainty about whether the fish has actually contacted or been caught by the lure.

The pattern is notable because additional metadata does not simply reduce hallucination. Instead, it is associated with more decisive event-level claims. In the image-only condition, none of the seven models explicitly states that the fish is hooked. In the metadata-guided condition, four models do so explicitly, and six of seven either state or strongly imply contact between the fish and the lure.

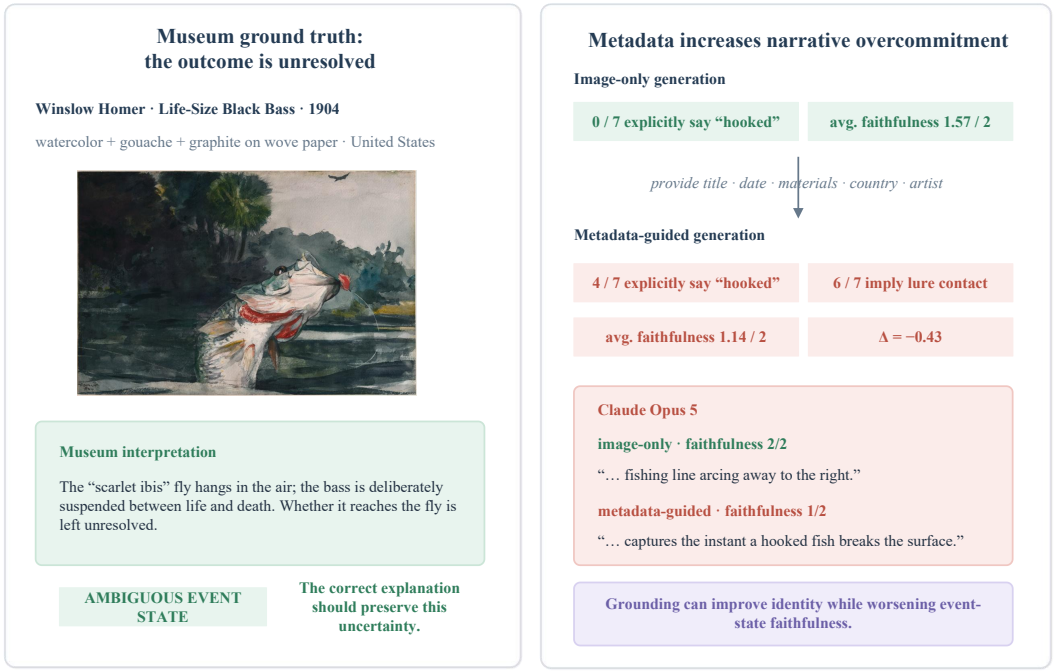


Figure 29: **Metadata-induced narrative overcommitment for `aic:189152`.** Left: the museum ground truth preserves an unresolved event state in which the fly remains in the air and the outcome is uncertain. Right: image-only generation never explicitly says that the fish is "hooked" (0/7) and achieves an average faithfulness of 1.57/2. After title, date, materials, country, and artist metadata are supplied, four of seven models explicitly say "hooked," six of seven claim or strongly imply lure contact, and average faithfulness drops to 1.14/2 ($\Delta = -0.43$).



Figure 30: **Paired image-only and metadata-guided responses for `aic:189152`.** Across several models, metadata-guided generation changes a cautious or descriptive account into a more causally committed narrative in which the fish is already hooked or physically pierced by the lure. The examples illustrate an event-state hallucination rather than an object-identity error.

At the same time, average faithfulness decreases from 1.57/2 to 1.14/2. This is the opposite of the provenance case above, where metadata grounding largely corrects the dominant error.

The paired excerpts in Figure 30 illustrate the shift at the model level. Claude Opus 5 changes from a relatively descriptive statement about a fishing line "arcing away to the right" to the definite claim that a "hooked fish breaks the surface." Gemini 3.1 Pro similarly moves from a fish that "strikes at" a red fly to a fly "hooked in its jaw." Kimi K2.6 changes from describing the mouth as open around a red lure to asserting that the fish is at "the precise moment of being hooked." MiniMax M3 is even more specific, introducing a "trailing hook" that "pierces its jaw." These examples show how a model can replace an unresolved depicted relation with an explicit causal event that the visual evidence and museum interpretation do not establish.

A plausible explanation is *event-script completion*. Once the system recognizes a canonical sport-fishing scene and receives authoritative metadata about the work, it may become more confident

in a familiar causal script: a fish approaches a lure, strikes it, becomes hooked, and struggles at the surface. The observed responses are consistent with this mechanism, although the present case study does not by itself establish the internal cause of the behavior. What can be established from the outputs is the evaluation failure: the generated explanation collapses an unresolved visual state into a definite event and, in several cases, adds unsupported physical details.

This case highlights a complementary notion of faithfulness. A model should not only avoid asserting incorrect object metadata; it should also preserve *epistemic uncertainty* when the depicted event is intentionally or visually unresolved.

Cross-case implication. The two cases expose different pathways to unfaithful museum explanation. For G-184, shared iconography leads models toward the wrong cultural prototype, after which provenance errors propagate into unsupported historical and material claims. For *Life-Size Black Bass*, the object and scene are broadly recognized, but models over-resolve an ambiguous event into a definite causal narrative. Metadata helps in the first case but worsens event-state faithfulness in the second. Consequently, a robust evaluation of MLLM museum interpretation should measure not only structured identification accuracy, but also cross-task consistency, claim-level factuality, preservation of uncertainty, and the effect of grounding on downstream narrative content.

L Web-Search Evaluation

We conduct the supplementary web-search experiment on *MuseFACET-testmini*, the 1,000-object subset of the full benchmark introduced in the main paper. All three evaluation settings are included: Identification, image-only Interpretation, and metadata-guided Interpretation. The experiment compares models with and without access to their native web-search functionality under otherwise identical task definitions, input settings, output formats, and evaluation criteria.

L.1 Web-Search Setup

We evaluate four proprietary models with their officially supported web-search functionality: Kimi K2.6, Qwen 3.7 Plus, Doubao Seed 2.0 Pro, and MiniMax M3. The corresponding search capabilities are accessed through their official services, including Alibaba Cloud Model Studio for Qwen, Volcengine Ark for Doubao, and the official Kimi and MiniMax services. We do not introduce an external retrieval system; instead, each model uses the native web-search functionality provided by its corresponding platform.

The benchmark task definitions, input settings, output formats, and evaluation criteria remain identical to those used in the corresponding no-search evaluation. The experiment therefore examines how access to external web information affects model performance on the same Identification and Interpretation tasks.

L.2 Complete No-Search Interpretation Results

The main paper reports the no-search Identification results on *MuseFACET-testmini* directly in Table 3. For completeness, we report here the corresponding no-search Interpretation results for all seven proprietary models. Tables 33 and 34 present the image-only and metadata-guided settings, respectively.

Table 33: No-search image-only Interpretation results for the seven proprietary models on MuseFACET-testmini. Each dimension is scored from 0 to 2, and *Total* is their sum.

Model	Total	Faith.	Comp.	Conc.	Depth	Acuity
Claude Opus 5	5.128	1.006	1.164	0.488	1.713	0.757
Gemini 3.1 Pro Preview	5.385	1.137	1.133	0.921	1.434	0.760
Qwen 3.7 Plus	3.886	0.597	0.794	0.919	1.182	0.394
GLM-5V Turbo	3.866	0.572	0.819	0.854	1.248	0.374
Doubao Seed 2.0 Pro	4.011	0.599	1.038	0.488	1.384	0.502
Kimi K2.6	4.289	0.668	0.965	0.575	1.541	0.539
MiniMax M3	3.715	0.703	0.885	0.660	1.100	0.367

Table 34: No-search metadata-guided Interpretation results for the seven proprietary models on MuseFACET-testmini. Each dimension is scored from 0 to 2, and *Total* is their sum.

Model	Total	Faith.	Comp.	Conc.	Depth	Acuity
Claude Opus 5	7.633	1.556	1.741	0.906	1.983	1.446
Gemini 3.1 Pro Preview	7.003	1.652	1.521	0.888	1.671	1.272
Qwen 3.7 Plus	6.419	1.281	1.232	1.371	1.511	1.024
GLM-5V Turbo	6.398	1.254	1.272	1.226	1.613	1.033
Doubao Seed 2.0 Pro	6.277	1.204	1.405	0.865	1.736	1.067
Kimi K2.6	6.353	1.230	1.339	0.906	1.837	1.040
MiniMax M3	6.126	1.178	1.366	1.014	1.559	1.010

L.3 Effect of Web Search

The complete search-enabled results and their changes relative to the corresponding no-search results are reported in Table 5 in the main paper.

For Identification, the effect of web search remains mixed. The strict *All* score increases by 1.00 percentage point for Kimi K2.6, 1.90 points for Doubao Seed 2.0 Pro, and 3.89 points for MiniMax M3, while Qwen 3.7 Plus decreases by 0.37 points. At the field level, changes are also uneven, and Materials accuracy decreases for all four models.

For image-only Interpretation, web search improves the Total score for Qwen 3.7 Plus and Doubao Seed 2.0 Pro by 0.440 and 0.613 points, respectively, but decreases it for Kimi K2.6 and MiniMax M3 by 0.247 and 0.192 points. Completeness and Depth increase for all four models, but these gains can be offset by reductions in Faithfulness or Conciseness.

The effect becomes more consistent under metadata guidance. All four models obtain higher Total scores with web search: +0.124 for Kimi K2.6, +0.324 for Qwen 3.7 Plus, +0.216 for Doubao Seed 2.0 Pro, and +0.384 for MiniMax M3. Completeness and Depth increase for all four models, whereas Conciseness decreases across all four. These results suggest that external retrieval is more useful when reliable object metadata already provides an identification anchor. Without such grounding, retrieved information may introduce irrelevant or conflicting evidence that acts as noise and interferes with the model's reasoning about the target object (Hong et al., 2024; Amiraz et al., 2025).

M Human Evaluation and LLM-Judge Validation

To validate whether the automatic evaluator is aligned with human judgment, we conduct a human evaluation of model-generated Interpretations. The human study uses the same five-dimensional rubric as the automatic evaluator: Faithfulness, Completeness, Conciseness, Depth, and Acuity, with each dimension scored independently from 0 to 2. The museum-provided description is treated as supporting factual, contextual, and interpretive evidence rather than as a single gold response, following the same principle used in the automatic evaluation.



Figure 31: Web-based interface used for human evaluation. The upper panel presents the museum-object image, normalized metadata, museum-provided reference description, and anonymized model-generated description. The lower panel presents the five evaluation dimensions—Faithfulness, Completeness, Conciseness, Depth, and Acuity—together with their dimension-specific scoring criteria and 0–2 rating options.

M.1 Evaluation Sample

We sample 200 model responses from the benchmark outputs using stratified random sampling. The sampling procedure considers object category, cultural region, and model configuration in order to avoid concentrating the human evaluation on a small subset of objects or model behaviors. The selected responses cover outputs from all seven proprietary models and both Interpretation settings evaluated in the main experiments.

Each sampled response is evaluated independently by all three human annotators. Model identities are hidden from the annotators, and the responses are presented in randomized order to reduce potential ordering and model-specific biases.

M.2 Annotators and Interface

The human evaluation is conducted by three annotators with university-level education and fluent English proficiency. Before the formal evaluation, the annotators receive training on the five evaluation dimensions and review the scoring guidelines and representative examples to establish a shared understanding of the rubric.

Evaluation is performed through a web-based interface. For each response, the interface presents the museum-object image, the normalized metadata, the museum-provided description, and the candidate model response. These materials provide the same core evidence used by the automatic evaluator while allowing annotators to inspect the image and contextual information jointly. The interface also presents the complete five-dimensional scoring rubric, allowing annotators to assign an independent 0–2 score for each evaluation dimension. Figure 31 shows the interface used in the formal human evaluation.

M.3 Agreement Analysis

We evaluate agreement at two levels. First, we measure consistency among the three human annotators using ordinal Krippendorff's α (Hayes & Krippendorff, 2007). Agreement is computed separately for Faithfulness, Completeness, Conciseness, Depth, and Acuity, as well as for the re-

Table 35: Detailed human evaluation agreement. H–H denotes inter-annotator agreement measured by Krippendorff's α, and H–LLM denotes Spearman's ρ between automatic scores and the mean of three human ratings. Human and LLM Avg. report the corresponding mean scores.

Dimension	H–H α	H–LLM ρ	Human Avg.	LLM Avg.
Total	**0.833**	**0.758**	**6.067**	**5.705**
Faithfulness	0.761	0.679	1.258	1.080
Completeness	0.685	0.622	1.280	1.265
Conciseness	0.666	0.546	1.287	1.010
Depth	0.648	0.531	1.353	1.600
Acuity	0.662	0.641	0.888	0.750

sulting 0–10 Total score. As reported in Table 35, the five individual dimensions obtain α values between 0.648 and 0.761, while the Total score reaches 0.833.

Second, we measure agreement between the automatic evaluator and human judgment. For each response and evaluation dimension, the three human ratings are averaged and compared with the corresponding GPT-5.5 score using Spearman's ρ. Human–LLM correlations range from 0.531 to 0.679 across the five dimensions, with a correlation of 0.758 for the Total score. These results indicate that the automatic evaluator broadly follows human scoring patterns across the proposed evaluation dimensions, with stronger agreement for the aggregate score.

The dimension-level results also reveal several consistent scoring tendencies. For Completeness, human and automatic mean scores are nearly identical (1.280 vs. 1.265). For Conciseness, the automatic evaluator assigns somewhat lower scores on average than human annotators (1.010 vs. 1.287), whereas for Depth it assigns somewhat higher scores (1.600 vs. 1.353). This suggests that the evaluator applies a relatively stricter criterion to conciseness while giving somewhat more credit to contextual or interpretive elaboration. These tendencies help explain the comparatively lower Human–LLM correlations for Conciseness and Depth, while the stronger agreement on the Total score indicates that the overall evaluation remains comparatively stable when the five dimensions are considered jointly.

N Cultural-Region Analysis

We compare relative model performance across six continent-level cultural regions: Europe, Asia, North America, Africa, South America, and Oceania. Because Identification and Interpretation use different score ranges, we normalize regional performance separately for each model and task. Identification is represented by the number of correctly predicted fields (0–4), while Interpretation uses the five-dimensional Total score (0–10).

Let $\mathcal{D}_c$ denote the set of objects assigned to region c, and let $s_t(m, x)$ denote the score of model m on object x for task t. We first compute the mean performance of model m on task t within region c as

$$\mu_{m,t,c} = \frac{1}{|\mathcal{D}_c|} \sum_{x \in \mathcal{D}_c} s_t(m, x).$$

For each model–task pair, we then compute the minimum and maximum regional means across the six regions,

$$\mu^{\min}_{m,t} = \min_{c \in \mathcal{C}} \mu_{m,t,c}, \qquad \mu^{\max}_{m,t} = \max_{c \in \mathcal{C}} \mu_{m,t,c},$$

where $\mathcal{C}$ denotes the set of six cultural regions. The relative regional score is defined as

$$r_{m,t,c} = 2\frac{\mu_{m,t,c} - \mu^{\min}_{m,t}}{\mu^{\max}_{m,t} - \mu^{\min}_{m,t}} - 1.$$

Thus, $r_{m,t,c} \in [-1, 1]$, where -1 and $+1$ denote the weakest and strongest regions, respectively, for the same model and task. If $\mu_{m,t}^{\max} = \mu_{m,t}^{\min}$, we set $r_{m,t,c} = 0$ for all regions. This normalization is used only to characterize relative regional variation within a model and task; it does not represent absolute performance differences across models or tasks.