

CARTBench: A Museum-Grounded Benchmark for Evidence-Faithful Chinese Art Understanding

Xuefeng Wei¹, Wang Zhixuan², Xuan Zhou¹, Zhi Qu¹, Hongyao Li²,
Yusuke Sakai¹, Hidetaka Kamigaito¹, Taro Watanabe¹

¹Nara Institute of Science and Technology (NAIST), Japan

²Liaoning Normal University, China

{xuefeng.wei.yb1, xuan.zhou.xc7}@naist.ac.jp

{wang, lihongyao1989}@lnnu.edu.cn

{qu.zhi.pv5, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

Abstract

We introduce CARTBENCH, a museum-grounded benchmark for evaluating vision-language models (VLMs) on Chinese artworks beyond short-form recognition and QA. CARTBENCH comprises four subtasks: CURATORQA for evidence-grounded recognition and reasoning, CATALOGCAPTION for structured four-section expert-style appreciation, REINTERPRET for defensible reinterpretation with expert ratings, and CONNOISSEURPAIRS for original-copy discrimination under visually similar confounds. CARTBENCH is built by aligning image-bearing Palace Museum objects from Wikidata with authoritative catalog pages, spanning five art categories across multiple dynasties. Across evaluated VLMs, high overall CURATORQA accuracy can mask sharp drops on evidence-heavy and style-to-period questions; long-form appreciation remains far from expert references; and original-copy discrimination remains near chance. A held-out protocol-replication pilot on 120 artworks from five external institutions reproduces the main task-difficulty pattern, while not preserving the exact overall model ranking.

1 Introduction

Vision-Language Models (VLMs) are increasingly deployed as general-purpose multimodal assistants. Generic VQA and shortcut-learning studies show that strong aggregate performance can coexist with reliance on dataset priors or non-robust cues (Agrawal et al., 2018; Geirhos et al., 2020). Separately, culturally grounded studies and benchmarks reveal language- or region-dependent gaps in image understanding (Ananthram et al., 2025; Nayak et al., 2024; Wang et al., 2025). A key capability nevertheless remains under-tested: expert-facing interpretation that is culturally grounded and explicitly supported by visual evidence. This matters for cultural heritage access and documentation, where culturally misaligned or unsupported inter-

pretations can reduce reliability and downstream usability.

Chinese art makes this evaluation gap particularly salient. Many visual conventions are period sensitive, and curator-facing understanding often requires linking observable cues, such as motifs, brushwork, seals, material appearance, and format, to coarse historical context rather than relying on surface recognition alone. We operationalize this requirement in our benchmark through knowledge-involved questions and style-to-period inference, where our experiments reveal consistent degradations even for strong models. original-copy discrimination is another challenging setting, and recent work has explored computational assistance for Chinese-painting authentication (Chen et al., 2025). We treat this task strictly as a diagnostic benchmark setting, not as a real-world authentication protocol, and find that current VLMs remain near chance on the 44-pair diagnostic set. Together, these observations motivate a benchmark that measures not only short-form correctness, but also evidence linkage, structured appreciation, defensible reinterpretation, and confound-aware original-copy discrimination.

We introduce CARTBENCH, a museum-grounded benchmark that targets curator-facing Chinese artwork understanding in a unified suite (see Figure 1).¹ CARTBENCH is built by aligning image-bearing Palace Museum objects from Wikidata (Vrandečić and Krötzsch, 2014) with authoritative catalog pages, followed by expert-guided curation and controlled task instantiation. It comprises four tasks corresponding to the above evaluation goals: (i) CURATORQA for evidence-grounded recognition and reasoning (14,421 questions over 1,589 artworks), (ii) CATALOGCAPTION for structured four-section appreciation (86 artworks), (iii) REINTERPRET for de-

¹Project repository: <https://github.com/Big-Sid/CARTBENCH-Chinese-Artwork-Benchmark>.

fensible reinterpretation with two-stage human ratings (25 canonical works), and (iv) CONNOISSEURPAIRS for original-copy discrimination on visually similar pairs (44-pair diagnostic set). Across representative VLMs, strong models achieve high overall CURATORQA accuracy but drop sharply on style-to-period inference; long-form appreciation diverges from expert references; and original-copy discrimination under close confounds remains near chance. We organize the study around three research questions: **RQ1:** To what extent can current VLMs move beyond short-form recognition to perform evidence-linked QA, structured expert-style appreciation, defensible reinterpretation, and visually confounded original-copy discrimination for Chinese art? **RQ2:** Which expert-critical conditions remain difficult despite strong aggregate performance, particularly knowledge-involved reasoning, style-to-period inference, evidence-faithful long-form generation, and plausibility under reinterpretation? **RQ3:** How reliable are the proposed data and evaluation protocols, and to what extent do the principal diagnostic patterns persist across model families and museum sources?

In this study, our contributions are as follows:

- We construct CARTBENCH, a museum-grounded benchmark for Chinese art spanning evidence-grounded QA, structured appreciation, defensible reinterpretation, and original-copy discrimination. It supports multi-period evaluation across five art categories while keeping high-stakes original-copy judgments explicitly scoped as diagnostic.
- We provide unified evaluation protocols combining automatic metrics, format compliance checks, expert-designed human rubrics, and data-quality controls. These protocols assess not only surface correctness, but also evidence linkage, structured long-form compliance, annotation reliability, and susceptibility to language-only shortcuts.
- We benchmark representative VLMs and identify systematic failure modes in evidence linking, style-to-period inference, long-form expert appreciation, defensible interpretation, and confound-aware original-copy discrimination. We find that high overall CURATORQA accuracy can coexist with substantial degradations on expert-critical conditions, which are obscured by short-form evaluation.

2 Related Work

vision-language evaluation and cultural grounding. Large-scale VQA and captioning benchmarks established standardized pipelines but primarily target generic recognition and short-form responses (Antol et al., 2015; Lin et al., 2014). Broader VLM evaluations such as MME and MM-Vet assess integrated multimodal capabilities (Fu et al., 2025; Yu et al., 2024), while hallucination studies show that model outputs can remain weakly grounded in visual evidence (Li et al., 2023). Recent culturally grounded benchmarks broaden culture-specific evaluation: CulturalVQA focuses on cultural visual question answering, while CVLUE provides a broader Chinese vision-language evaluation suite (Nayak et al., 2024; Wang et al., 2025). Cross-lingual artwork explanation studies likewise reveal language-dependent gaps in explanation quality (Ozaki et al., 2025).

Art- and museum-centered vision-language resources. Art-specific datasets explore dimensions beyond object naming: ArtEmis pairs artworks with viewer emotions and grounded explanations, while ArtELingo-28 extends affective artwork captioning across 28 languages (Achlioptas et al., 2021; Mohamed et al., 2024). Museum-centered QA resources such as MuseumQA emphasize artifact knowledge and fine-grained factual questioning in curated settings, while recent systems for Chinese painting and calligraphy integrate structured resources and external documents (Jin et al., 2024; Wan et al., 2025). Table 1 provides a compact task-coverage comparison with representative related resources. These resources are complementary rather than inferior alternatives; CARTBENCH differs in jointly covering these task settings rather than claiming comprehensive cultural or museum coverage.

Connoisseurship and computational assistance. Authentication and connoisseurship have attracted increasing attention as high-stakes cultural heritage applications. ACPAS proposes an expert-assistance system for authenticating ancient Chinese paintings via LLM-based agents, and ArtEyer develops a GPT-based authentication agent augmented with contextual data visualizations (Chen et al., 2025; Tang et al., 2024). While these systems demonstrate the feasibility of computational support for connoisseurship, they are not standardized VLM benchmarks with controlled confounds and consis-

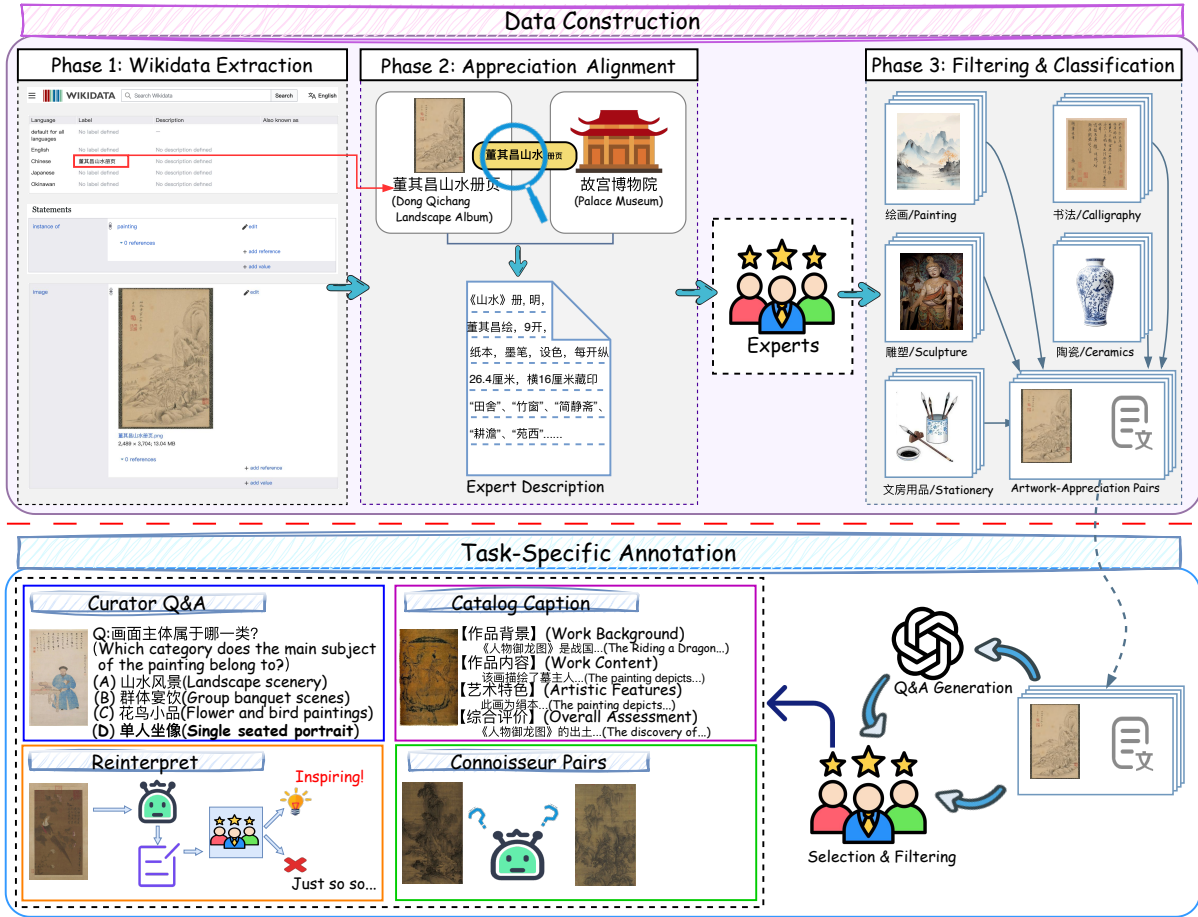


Figure 1: Overview of CARTBENCH construction and task instantiation. **Top: Phase 1** retrieves image-bearing Palace Museum objects from Wikidata, **Phase 2** aligns them to official catalog pages to collect curatorial descriptions, and **Phase 3** performs expert filtering and category assignment to yield museum-grounded artwork-appreciation pairs. Bottom: the curated pairs are instantiated into four tasks: CURATORQA (evidence grounding), CATALOGCAPTION (structured appreciation), REINTERPRET (expert-rated reinterpretation), and CONNOISSEURPAIRS (confound-aware original-copy discrimination).

tent scoring protocols.

3 CARTBENCH: Data Construction

Motivated by the gaps identified above, we construct CARTBENCH through a museum-grounded pipeline that aligns Wikidata entities with authoritative catalog pages, followed by expert-guided curation and controlled annotation. The high-level pipeline is reusable, but its source alignment, evidence definitions, annotation constraints, and human rubrics are tailored here to Chinese art. Figure 1 overviews the museum-grounded construction pipeline used to build CARTBENCH. From the resulting museum-grounded artwork-description pairs, four subtasks are instantiated to evaluate complementary aspects of Chinese artwork understanding, as outlined below.

3.1 Alignment via Wikidata

First, we retrieve museum objects with images using Wikidata Query Service (Vrandečić and Krötzsch, 2014). We restrict to objects associated with the Palace Museum (Beijing)² and require image availability. This initial Wikidata query yields 127,601 image-bearing items linked to the Palace Museum. We then align each object by title to the museum catalog page and collect the on-page curatorial description.

3.2 Category and dynasty coverage

Following expert feedback and Wikidata metadata, we assign each item to one of five categories: painting, calligraphy, sculpture, ceramics, and stationery. The collection spans multiple dynasties rather than concentrating on a single period. Figure 2 summa-

²<https://intl.dpm.org.cn/index.html?l=en>

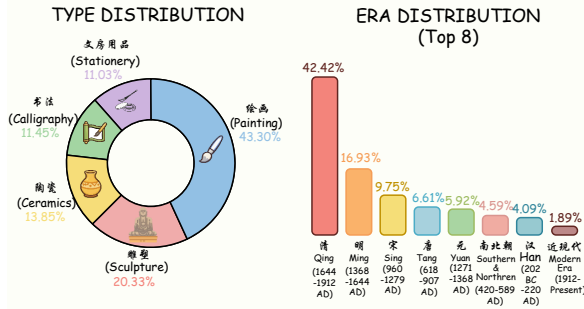


Figure 2: Distributions of CURATORQA entries: (left) five art type categories; (right) top-8 merged eras.

rizizes the resulting type and era distributions. For reference, Appendix A.19 lists the dynasties and approximate date ranges used in our era bins.

3.3 Task-specific annotation

We build four tasks with complementary supervision and evaluation. At a high level, the tasks target four distinct aspects of curator-facing Chinese artwork understanding. CURATORQA focuses on evidence-linked recognition and reasoning under controlled answer options; CATALOGCAPTION evaluates structured expert-style appreciation in a fixed four-section format; REINTERPRET probes whether models can produce novel yet defensible readings that remain coherent and image-grounded under expert judgment; and CONNOISSEURPAIRS tests original-copy discrimination on visually similar pairs. The bottom of Figure 1 summarizes how the curated pairs are instantiated into four tasks. We detail each task’s format and annotation below. The evaluation protocols are provided in Section 4.

CURATORQA. Task format. Each instance pairs an artwork image with a question; the model selects from candidate options for multiple-choice items or returns a binary answer for verification items. CURATORQA evaluates whether VLMs can answer curator-style questions using explicit visual evidence and relevant art-domain knowledge. We generate QA pairs from aligned artwork images and authoritative museum catalog descriptions using gpt-5.2 under controlled prompting, which also produces plausible distractors, P1/P2 difficulty labels, and QA-type categories following a fixed schema. A stratified expert audit of 1,000 questions finds one erroneous item (0.1%). Under an exact-binomial calculation, the one-sided 95% upper bound is 0.47%; because the audit uses stratification and an artwork-level cap, we treat this value as an audit diagnostic rather than a design-

Resource	Mus.	Struct.	Reint.	O/C
VQA/MME/MM-Vet	—	—	—	—
CulturalVQA/CVLUE	—	—	—	—
ArtEmis/ArtELingo-28	—	—	—	—
MuseumQA	✓	—	—	—
CARTBENCH	✓	✓	✓	✓

Table 1: Compact task-coverage comparison. Mus.=museum/artifact setting; Struct.=structured appreciation; Reint.=expert-rated reinterpretation; O/C=original-copy diagnostic.

based confidence bound for the full corpus (Appendix A.1). After automated constraint filtering and this targeted expert audit, the final set contains 14,421 questions over 1,589 artwork entries. P1 questions are answerable from visible evidence alone, whereas P2 questions require combining visual cues with art knowledge, such as inferring dynasty from style. Questions cover six QA types, including subject recognition, scene/activity classification, composition or format, visible technique or style cues, iconography detection, and style-to-period inference. Appendix A.18 reports the main statistics and distributions.

CATALOGCAPTION. Task format. Given an artwork image, the model generates an expert-style appreciation in four fixed sections: Background, Content, Artistic Characteristics, and Overall Evaluation. We construct an expert-grounded set of 86 artworks via multi-stage curation: specialists collect candidates from professional Chinese art books and catalogues, we enforce Wikidata traceability to ensure each work has a stable image, and we filter low-quality images and non-informative texts. Experts then check and rewrite the reference appreciation into the same four-section schema, and we release both a full reference and section-wise references for holistic and per-dimension evaluation. These references are standardized expert-curated targets rather than raw catalogue prose; this standardization reduces stylistic variance and enables automatic scoring, format compliance checks, and rubric-based expert validation (Appendix A.4).

REINTERPRET. Task format. Given an artwork image, the model produces an open-ended reinterpretation that goes beyond literal description while remaining image-grounded and culturally appropriate. To enable a controlled assessment of interpretive novelty, art specialists selected 25 canonical Chinese artworks with well-established art-historical interpretations. We use them as anchors

Model	Acc.	QA type Acc.						Category Acc.				
		QA1	QA2	QA3	QA4	QA5	QA6	Callig.	Station.	Paint.	Ceram.	Sculpt.
Qwen family												
Qwen3-VL-235B	0.84	0.88	0.94	0.88	0.89	0.56	0.81	0.85	0.78	0.87	0.79	0.84
Qwen3-VL-30B	0.80	0.85	0.91	0.85	0.88	0.42	0.78	0.83	0.74	0.83	0.76	0.80
Qwen2.5-VL-72B	0.81	0.85	0.93	0.85	0.87	0.53	0.76	0.83	0.74	0.85	0.75	0.80
Qwen2.5-VL-32B	0.80	0.84	0.91	0.83	0.87	0.53	0.74	0.82	0.76	0.83	0.73	0.78
Qwen2.5-VL-7B	0.76	0.82	0.89	0.78	0.81	0.44	0.72	0.75	0.70	0.80	0.68	0.75
GLM family												
GLM-4.5V	0.57	0.59	0.67	0.62	0.62	0.29	0.56	0.59	0.52	0.60	0.54	0.57
GPT family												
gpt-5-mini	0.79	0.84	0.90	0.83	0.86	0.46	0.78	0.83	0.74	0.84	0.72	0.78
gpt-5-nano	0.59	0.62	0.73	0.53	0.65	0.27	0.61	0.59	0.55	0.63	0.53	0.56
Gemini family												
gemini-2.5-flash	0.48	0.51	0.54	0.50	0.52	0.31	0.47	0.48	0.45	0.51	0.45	0.48

Table 2: CURATORQA accuracy breakdown (overall, by question type, and by art category). We report exact-match accuracy on 14,421 questions. Category abbreviations: Callig.=calligraphy, Station.=stationery, Paint.=painting, Ceram.=ceramics, Sculpt.=sculpture.

to test whether model interpretations can go beyond conventional readings while remaining grounded in the image and culturally appropriate.

CONNOISSEURPAIRS. Task format. Each instance presents an original-copy or original-imitation pair; the model selects which side corresponds to the benchmark-designated original. We use this task only as a controlled diagnostic benchmark setting, not as a real-world authentication, provenance, certification, or valuation protocol. To test this capability, art specialists curated 44 visually similar pairs, prioritizing close visual similarity, matched subject matter and medium where possible, stable label documentation and expert screening, and screening against trivial cues such as obvious cropping artifacts or heavy watermarks.

4 Evaluation Protocols

4.1 CURATORQA

We evaluate exact-match accuracy under constrained decoding.

4.2 CATALOGCAPTION

We evaluate CATALOGCAPTION as a long-form, expert-style appreciation generation task with both automatic similarity metrics and structured format compliance. To avoid confounds from Chinese word segmentation, we compute overlap-based metrics using *character-level* tokenization, including ROUGE-L (Lin, 2004), BLEU-4 (Papineni et al., 2002), and a CIDEr-like consensus similarity metric (Vedantam et al.,

2015). We also report BERTScore (F1) (Zhang et al., 2020) using bert-base-chinese with rescale_with_baseline=True. As an auxiliary semantic signal beyond lexical overlap, we compute embedding cosine similarity (EmbSim) with shibing624/text2vec-base-chinese. Because CATALOGCAPTION targets a fixed four-section schema, we additionally measure *Header rate*, the fraction of outputs that contain all four required section headers. Finally, we aggregate the core automatic metrics into a single KPI score:

$$\text{KPI} = \lambda_1 \text{BERTScore}_{F1} + \lambda_2 \text{CIDEr-like} + \lambda_3 \text{ROUGE-L} + \lambda_4 \text{BLEU-4}, \quad (1)$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^4 \lambda_i = 1$. Unless otherwise specified, we use fixed weights $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{0.45, 0.25, 0.20, 0.10\}$. We assign higher weights to semantic metrics (e.g., BERTScore) to better capture the interpretive nuance of appreciation text. We further verify in Appendix A.3 that our model rankings remain stable under diverse weight perturbations. Because automatic metrics are only scalable proxies for open-ended appreciation, we additionally report a rubric-based expert evaluation in Appendix A.4. The expert study broadly corroborates the automatic results: Qwen variants receive higher rubric scores than Gemini on evidence-grounded interpretation, cultural/art-historical appropriateness, and insight depth, while Qwen3-VL-235B shows the lowest major factual/hallucination error rate among the three evaluated models.

4.3 REINTERPRET

We evaluate REINTERPRET with a custom two-stage questionnaire designed for Chinese-art reinterpretation. Stage 1 performs a plausibility gate, filtering outputs with major misreadings, clear violations of widely accepted art-historical consensus, or unsupported factual fabrications. Stage 2 rates the remaining outputs on a 1-5 Likert scale along five criteria: **Interpretive Novelty, Integrative Coherence, Evidence-based Reasoning, Elaboration & Expressiveness, and Perceived Creative Insight**. The full questionnaire is provided in Appendix A.14. Appendix A.6 further reports inter-annotator agreement, showing moderate-to-substantial agreement for the Stage 1 plausibility gate and reasonable-to-high reliability for three-rater averaged Stage 2 scores.

4.4 CONNOISSEURPAIRS

We evaluate CONNOISSEURPAIRS using pairwise accuracy over the 44 original-copy pairs. Because this diagnostic set remains small, we report accuracy and avoid fine-grained leaderboard-style claims. To better understand failure mechanisms beyond a single accuracy number, we additionally conduct a qualitative error analysis with an expert-facing questionnaire; the full instrument is provided in Appendix A.15.

5 Experimental Setup

We benchmark representative vision-language models from the open-weight Qwen2.5-VL and Qwen3-VL families (Bai et al., 2025b,a) and the GLM-V family (Team et al., 2026), together with proprietary models from OpenAI (gpt-5-mini and gpt-5-nano for evaluation, and gpt-5.2 for LLM-assisted question generation)³ and Google (gemini-2.5-flash)⁴. Qwen and GLM models are accessed through the AI Ping API gateway; exact access methods and model identifiers are reported in Appendix A.17. We audited publicly documented Chinese-specialized candidates against the full-benchmark requirements and found no strictly Chinese-monolingual model compatible with all four tasks and represented art forms. We therefore omit an incomparable monolingual baseline; details are in Appendix A.16.

³OpenAI Models documentation: <https://platform.openai.com/docs/models>; accessed 2026-05-03.

⁴Google Gemini model documentation: <https://ai.google.dev/gemini-api/docs/models>; accessed 2026-05-03.

Subset	mean Δ (CI ₉₅)	P _{perm}
QA1	0.0426 [0.0279, 0.0574]	< 0.05
QA2	0.0354 [0.0215, 0.0485]	< 0.05
QA3	0.0435 [0.0301, 0.0568]	< 0.05
QA4	0.0345 [0.0224, 0.0462]	< 0.05
QA5	0.1022 [0.0735, 0.1302]	< 0.05
QA6	0.0295 [0.0169, 0.0423]	< 0.05
P1	0.0381 [0.0312, 0.0449]	< 0.05
P2	0.0626 [0.0488, 0.0766]	< 0.05
Calligraphy	0.0254 [0.0061, 0.0435]	< 0.05
Painting	0.0294 [0.0199, 0.0385]	< 0.05
Sculpture	0.0678 [0.0522, 0.0827]	< 0.05
Ceramics	0.0688 [0.0508, 0.0883]	< 0.05
Stationery	0.0445 [0.0266, 0.0620]	< 0.05

Table 3: Subset-wise robustness of the Qwen3-VL-235B-GPT-5-mini difference on CURATORQA. Here, Δ denotes the equal-artwork-weight mean accuracy difference on answered items; 95% confidence intervals use artwork-clustered bootstrap resampling, and p -values use artwork-clustered sign-flip permutation tests.

Weights scenario	Spearman	Kendall	Top-1
Uniform weights	0.991	0.964	1.000
Dirichlet ($\alpha=5$)	0.991	0.964	1.000
Dirichlet ($\alpha=1$)	0.964	0.891	1.000
Local Gaussian ($\sigma=0.05$)	0.991	0.964	1.000

Table 4: KPI weight sensitivity for CATALOGCAPTION. Spearman/Kendall are rank correlations with the default-weight ordering. For stochastic schemes (Dirichlet and Local Gaussian), we report the median over samples.

For each task, we use a unified prompt template and request temperature=0 and top_p=1 where supported, with one response per input. These settings reduce sampling variability but do not guarantee bitwise-deterministic outputs for models accessed through hosted APIs or gateways. We post-process outputs into task-specific formats. Full prompting templates, inference settings, and implementation details are in Appendix A.10-A.17.

6 Results and Analysis

6.1 CURATORQA

Overall performance. Table 2 summarizes the main CURATORQA results on an aligned subset of 14,421 questions for which predictions can be consistently compared across model runs. Among all evaluated models, Qwen3-VL-235B achieves the highest overall accuracy (0.84). GPT-5-mini achieves 0.79, while Gemini-2.5-flash is substantially lower at 0.48. Within the GLM family, GLM-4.5V obtains 0.57.

Model	KPI↑	BERTScore _{F1} ↑	CIDEr-like↑	ROUGE-L↑	BLEU-4↑	EmbSim↑	Header rate↑
Human baseline							
Human (mean of 3)	0.698	0.746	0.673	0.699	0.546	0.982	0.988
Qwen family							
Qwen2.5-VL-72B	0.261	0.329	0.205	0.237	0.146	0.876	1.000
Qwen3-VL-235B	0.258	0.319	0.223	0.223	0.138	0.893	1.000
Qwen3-VL-30B	0.256	0.322	0.224	0.222	0.112	0.891	1.000
Qwen2.5-VL-7B	0.245	0.305	0.196	0.224	0.137	0.873	1.000
Qwen2.5-VL-32B	0.235	0.295	0.211	0.204	0.087	0.878	1.000
GLM family							
GLM-4.5V	0.245	0.315	0.194	0.217	0.113	0.878	0.988
GLM-4.6V	0.241	0.302	0.209	0.211	0.100	0.882	0.977
GPT family							
gpt-5-nano	0.225	0.276	0.193	0.205	0.113	0.861	1.000
gpt-5-mini	0.224	0.275	0.198	0.202	0.103	0.871	1.000
Gemini family							
gemini-2.5-flash	0.170	0.201	0.160	0.160	0.069	0.812	0.256

Table 5: CATALOGCAPTION evaluation on 86 artworks. Overlap-based metrics (ROUGE-L / BLEU-4 / CIDEr-like) use character-level tokenization. BERTScore_{F1} uses bert-base-chinese with rescale_with_baseline=True; EmbSim is cosine similarity under text2vec-base-chinese. Header rate measures whether all four required section headers appear in the output.

Model	Gate pass↑	Primary avg (D1-D3)↑	All avg (D1-D5)↑
gemini-2.5-flash	0.707 [0.600, 0.813]	3.673 [3.476, 3.887]	3.723 [3.528, 3.933]
GLM-4.6V	0.853 [0.760, 0.933]	3.649 [3.522, 3.784]	3.679 [3.575, 3.796]
gpt-5-mini	0.867 [0.787, 0.947]	3.567 [3.480, 3.656]	3.617 [3.529, 3.708]
qwen3-vl-235B	0.880 [0.787, 0.960]	3.747 [3.622, 3.878]	3.728 [3.621, 3.836]

Table 6: REINTERPRET human creativity evaluation on 25 artworks with three expert raters. Gate pass is the Stage 1 plausibility pass rate; Stage 2 uses 1-5 Likert ratings and is computed only on passed outputs. We report artwork-level macro-averages with 95% bootstrap confidence intervals. Primary avg averages D1-D3, and All avg averages D1-D5.

Category and QA-type effects. To substantiate subset-level difficulty patterns beyond aggregate accuracy, we test cross-model consistency by computing per-artwork accuracies and applying a sign test across models. Category effects are systematic: for all evaluated models, *stationery* and *ceramics* are consistently harder than *painting* (median gaps -0.096 and -0.083 ; $p=0.002$) and harder than *calligraphy* (median gaps ≈ -0.071 ; $p=0.002$). In contrast, QA-type variation is dominated by knowledge-heavy inference: QA5 (style→period inference) shows the largest degradation across models (e.g., Qwen3-VL-235B: 0.56 on QA5 vs. 0.88-0.94 on QA1-QA4; GPT-5-mini: 0.46 vs. 0.83-0.90), aligning with QA5’s intent to probe nontrivial stylistic-to-historical mapping.

Task-conditional comparison. To characterize where two strong evaluated systems differ on CURATORQA, we compare Qwen3-VL-235B with GPT-5-mini on the same 14,421-question popula-

tion. Qwen3-VL-235B exceeds GPT-5-mini by +4.51 accuracy points overall, and the difference is statistically significant under both a paired McNemar test on matched questions and an artwork ID-clustered test. At the item level, the gap is +3.88 points on P1, +6.59 on P2, and +10.22 on QA5. Artwork-clustered contrasts further show stronger concentration in P2 than P1 (+2.72 points; 95% CI [+1.14, +4.31]) and in QA5 than non-QA5 (+6.34 points; [+3.61, +9.19]). Detailed paired and artwork-clustered analyses are provided in Appendix A.1. We interpret this comparison as task-conditional evidence between two named systems, rather than evidence about provider geography, cultural orientation, or training-data provenance.

Findings. Overall, CURATORQA yields high aggregate accuracy for strong models, yet exposes a consistent failure mode on style-to-period reasoning. The Qwen3-VL-235B-GPT-5-mini comparison further shows that their performance differ-

ence is concentrated more strongly in knowledge-involved and art-historical inference subsets, particularly P2 and QA5. This task-localized result does not identify the source of the difference in terms of model geography, cultural orientation, or training-data composition. Additional audits and controls (Appendix A.1) find one QA error in 1,000 audited items and large no-image/image-shuffle drops, supporting CURATORQA reliability and visual ground.

6.2 CATALOGCAPTION

Results overview. Table 5 reports CATALOGCAPTION results under the evaluation protocol in Section 4.2. Across all model families, automatic scores remain far below the human baseline, indicating that expert-style appreciation generation is still a major bottleneck for current VLMs even when they can follow the required format.

Cross-family comparison and interpretation of metrics. Within open-weight models, the top three KPI scores are achieved by Qwen variants, while GLM variants are competitive with the smaller Qwen models. The GPT baselines (gpt-5-mini/nano) remain below the top Qwen variants, while gemini-2.5-flash performs substantially worse. Interestingly, strong models reach relatively high embedding similarity ($\text{EmbSim} \approx 0.87\text{-}0.89$), yet overlap-based metrics remain low in absolute value (e.g., $\text{ROUGE-L} \approx 0.20\text{-}0.24$), suggesting that models often capture coarse semantic intent but fail to match the expert reference in structured, section-specific informational units and stylistic conventions required by CATALOGCAPTION.

Large gap to expert-written appreciation. Human writers achieve $\text{KPI} = 0.698$ on average (three independent annotations), whereas the best model (Qwen2.5-VL-72B) reaches $\text{KPI} = 0.261$. This corresponds to a 0.437 absolute KPI gap, with the best model attaining only $\sim 37\%$ of the human KPI. A similar trend holds across all component metrics: they substantially trail humans on BERTScore_{F1} , CIDEr-like, ROUGE-L, and BLEU-4, suggesting that generations are often incomplete in section-wise coverage and deviate from expert references in fine-grained, evidence-linked descriptions.

Robustness to KPI weighting. To ensure our conclusions do not hinge on a particular KPI weighting, we perturb metric weights using uniform reweighting, Dirichlet sampling, and local Gaussian perturbations (Appendix A.3). The top-

ranked model is unchanged across all tested scenarios (Top-1 same = 1.0), and rankings remain highly consistent with the default ordering (Spearman 0.964-0.991; Kendall 0.891-0.964; Table 4), supporting the robustness of the main findings.

Findings. CATALOGCAPTION shows that expert-style, long-form appreciation remains far below human performance. The strongest Qwen variants lead among evaluated models, yet automatic metrics indicate weak reference alignment despite high semantic similarity. The main conclusions are stable to KPI reweighting. A rubric-based expert study in Appendix A.4 corroborates this trend, with Qwen variants outperforming Gemini on evidence grounding, cultural appropriateness, and insight.

6.3 REINTERPRET

Results overview. Table 6 summarizes REINTERPRET human ratings on 25 canonical artworks with three raters. We report the Stage 1 plausibility gate pass rate and Stage 2 creativity scores on a 1-5 Likert scale. Following the questionnaire design (Appendix A.14), Stage 2 scores are computed only for gate-passed outputs; gate-failed cases are excluded from the mean. All scores are macro-averaged over artworks (equal weight per artwork), and 95% confidence intervals are obtained by bootstrap resampling over the 25 artworks.

Interpretation. Across models, Stage 2 scores on the passed subset fall into a narrow range (Primary avg: 3.57-3.75), indicating broadly similar perceived creativity once outputs meet the basic defensibility threshold. Under our anchored rubric (Appendix A.14), 3 corresponds to a competent, non-trivial reinterpretation, whereas 4 reflects a clearly distinctive and well-supported reading. In contrast, the Stage 1 gate provides a clearer separation: pass rates span 0.707-0.880, suggesting that the dominant bottleneck is not stylistic expression but avoiding errors that invalidate an otherwise plausible interpretation. Appendix A.5 reports the distribution of gate-failure types; Appendix A.6 reports inter-annotator agreement, supporting the reliability of averaged human judgments while cautioning against over-interpreting small Stage 2 differences.

Findings. Overall, REINTERPRET is constrained by defensibility. Models achieve higher effective creativity primarily by passing the plausibility gate more consistently, rather than by substantially im-

Model	Acc.	Model	Acc.
gpt-5-mini	0.545	Qwen2.5-VL-7B	0.477
Qwen2.5-VL-72B	0.545	Qwen3-VL-235B	0.477
Qwen2.5-VL-32B	0.523	GLM-4.5V	0.477
Qwen3-VL-30B	0.500	gemini-2.5-flash	0.477
gpt-5-nano	0.386		

Table 7: CONNOISSEURPAIRS accuracy on 44 original-copy pairs.

proving Stage 2 scores on already-valid outputs. In this setting, qwen3-vl-235B is the most reliable, while gemini-2.5-flash is most limited.

6.4 CONNOISSEURPAIRS

Table 7 reports original-copy accuracy on the 44-pair diagnostic set, where each model chooses which side corresponds to the benchmark-designated original.

Performance remains near chance. The highest observed accuracy is 24/44 (0.545; Wilson 95% CI [0.401, 0.683]), and the interval includes chance performance (0.5). Given the width of this interval, we do not interpret small differences among models as a stable ranking. The main takeaway is that strong general VLM performance does not transfer to visually close original-copy comparison: models appear unreliable at using global brush-and-ink coherence, style consistency, and copying artifacts under close visual confounds. Thus, we use CONNOISSEURPAIRS only as a diagnostic stress test. Appendix A.8 summarizes expert error attribution.

6.5 Cross-Institution Protocol Replication

To test whether the main diagnostic patterns are specific to the Palace Museum source, we conduct a held-out protocol-replication pilot on 120 Chinese artworks from five external institutions: the Smithsonian National Museum of Asian Art, the Victoria and Albert Museum, the Art Institute of Chicago, the Cleveland Museum of Art, and the Metropolitan Museum of Art. We preserve the original task definitions and evaluation protocols while changing the source institutions.

As shown in Table 8, the exact overall model ranking does not reproduce across institutions. Nevertheless, the principal diagnostic pattern transfers strongly: P2 is harder than P1 for every tested model, QA5 style-to-period inference remains the weakest QA type for every model, and the per-model Spearman correlation between the main and external QA-type profiles ranges from 0.81 to

Model	External	Main	Δ
GPT-5-mini	0.792	0.795	−0.003
Qwen2.5-VL-72B	0.789	0.808	−0.019
Qwen3-VL-30B	0.778	0.800	−0.022
Qwen3-VL-235B	0.777	0.840	−0.063
Qwen2.5-VL-7B	0.756	0.760	−0.004

Table 8: Cross-institution replication of CURATORQA. External accuracy is computed on 712 closed-form questions from 120 artworks across five external institutions. Main denotes the corresponding main-benchmark accuracy for the same five models. Δ is External−Main.

0.99. Four of the five overall accuracies are within 2.2 percentage points of the main result, whereas Qwen3-VL-235B decreases by 6.3 points.

We further replicate the two long-form protocols. On 24 external CATALOGCAPTION artworks, Qwen3-VL-235B, GPT-5-mini, and Qwen2.5-VL-7B obtain KPI scores of 0.289, 0.288, and 0.269, respectively, indicating that the two stronger systems are effectively tied under the automatic metric. On 12 external REINTERPRET artworks, Qwen3-VL-235B and GPT-5-mini obtain Stage 2 composites of 4.285 [4.103, 4.436] and 3.964 [3.861, 4.064], respectively. On the 11 artworks that majority-pass for both models, the paired Qwen-minus-GPT difference is +0.300, with an artwork-bootstrap 95% CI of [+0.091, +0.500]. Overall, the pilot supports initial transfer of the benchmark’s principal difficulty and failure patterns across museum sources, but not an institution-invariant model ranking. Therefore, we interpret CARTBENCH as a Palace Museum-grounded benchmark with initial cross-institution evidence rather than a universal evaluation standard for Chinese-art understanding. Full construction, audit, and evaluation details are provided in Appendix A.2.

7 Conclusion

We introduced CARTBENCH, a museum-grounded benchmark for evidence-faithful Chinese art understanding across four curator-facing tasks. Results show that high QA accuracy can mask persistent weaknesses in style-to-period inference, expert-style appreciation, defensible reinterpretation, and diagnostic original-copy comparison. Our audits and rubric-based evaluations support the benchmark’s reliability and highlight the need for VLMs that more effectively connect visual evidence to culturally specific concepts, uncertainty, and medium-specific style cues. A held-out five-institution replication transfers the task-difficulty profile.

Limitations

Diagnostic scope of original-copy discrimination. CONNOISSEURPAIRS contains 44 original-copy pairs. Because curating visually close and defensible pairs requires substantial expert effort, the task remains a diagnostic stress test rather than a stable leaderboard or any form of real-world authentication protocol. Accuracy estimates should therefore be interpreted with uncertainty and used for failure analysis rather than deployment claims.

LLM-assisted question generation may introduce residual artifacts. CURATORQA questions are LLM-assisted and then filtered with constraints and expert review. We add a 1,000-question expert audit, full-set option-bias diagnostics, and matched no-image/image-shuffle controls, but residual template regularities and language priors may remain. Future work should further incorporate human-authored adversarial or contrastive items.

Pretraining exposure cannot be excluded. The source museum pages and some canonical artworks are publicly available, and we cannot determine whether individual models encountered related images or descriptions during pretraining. The benchmark therefore evaluates task performance under realistic public-source exposure rather than guaranteeing contamination-free knowledge acquisition.

Cross-institution evidence remains a pilot. The main benchmark is grounded in the Palace Museum catalogue. Our held-out five-institution replication indicates that the task-difficulty profile and weakest QA type transfer, but the exact overall model ranking does not. The pilot is relatively small, relies mainly on English-language catalogues outside China, and covers only five institutions. It therefore provides initial protocol-transfer evidence rather than a culture- or institution-invariant generalization claim.

Automatic metrics only partially capture expert appreciation quality. For CATALOGCAPTION, overlap-based metrics and embedding similarity provide scalable signals but do not fully reflect whether an appreciation is truly expert-like, culturally grounded, and well supported by visual evidence. We therefore add rubric-based expert evaluation, while acknowledging that larger-scale expert preference judgments and evidence-unit evaluation remain valuable future extensions.

Ethical Considerations

This section summarizes potential risks and the measures we take to support responsible use of CARTBENCH.

Potential risks and misuse. This benchmark evaluates culturally grounded interpretation and original-copy discrimination for Chinese artworks. A key risk is overtrust: model outputs may be misused as expert authentication advice, leading to financial or reputational harm in real transactions. A second risk is dual use: insights from original-copy comparison could be exploited to improve high quality imitations. We mitigate these risks by positioning CONNOISSEURPAIRS as a diagnostic evaluation, requiring cue-based justifications, and emphasizing that outputs must not be used for authentication, valuation, provenance certification, market claims, or deployment decisions.

Resource release, licensing, and attribution. CARTBENCH draws on museum catalog pages, Wikidata, and expert-curated references from professional art books and catalogues. To support reproducibility, we will publicly release the evaluation code, benchmark metadata, prompts, and author-created derived annotations. The evaluation code will be released under the MIT License, and the author-created derived annotations under CC BY-NC 4.0. For source materials with license or attribution requirements, we provide stable identifiers, source URLs, and provenance information, and redistribute original images or catalog texts only when permitted by the corresponding licenses or rights holders. Wikidata content is available under CC0, while third-party museum materials remain subject to the corresponding source institutions' terms.

Use of AI assistants. AI assistants were used primarily for language editing and manuscript proof-reading during manuscript preparation, e.g., the camera-ready phase. All scientific content and final revisions were reviewed and approved by the authors.

Privacy and data protection. The benchmark is derived from museum object pages and is not intended to contain personally identifying information. We screen collected texts for common identifying patterns such as phone numbers, email addresses, and physical addresses, and remove any matches if found. Human ratings are stored in anonymized form and reported only in aggregate.

Cultural context and interpretive validity. Chinese art interpretation depends on culturally specific concepts and conventions, including medium and format terminology and style-related norms. Model generations may introduce conceptually misaligned claims or overconfident attributions that are not supported by visible evidence. We therefore emphasize evidence-linked description and require explicit uncertainty when key information cannot be reliably inferred from the image, and our human rubrics penalize unsupported factual assertions and culturally inappropriate interpretations.

Scope and accessibility. CArTBench is intentionally centered on Chinese art to broaden culturally specific VLM evaluation in an expert cultural-heritage domain. We formulate prompts and rubrics in Chinese to reflect the primary language and conventions of the underlying museum sources and real-world curatorial practice, rather than as a claim of culture-agnostic coverage. To support broader use, we provide English translations of prompts and documentation, and we view bilingual and cross-cultural extensions as an important direction for future work.

Acknowledgments

This research was supported by the China Scholarship Council (CSC) under grant numbers 202308070070 (Xuan Zhou) and 202308070072 (Xuefeng Wei), and was partly supported by JSPS Kakenhi Grant Number JP26H02537. Zhixuan Wang was supported by the Basic Scientific Research Project for Higher Education Institutions of Liaoning Province, administered by the Liaoning Provincial Department of Education, under the project “Evaluation and Enhancement of Multimodal Large Models for Understanding Chinese Calligraphy and Painting Cultural Heritage” (Grant No. LJ112610165002).

References

- Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. 2021. [Artemis: Affective language for visual art](#). In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11564–11574.
- Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. [Don’t just assume; look and answer: Overcoming priors for visual question answering](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.
- Amith Ananthram, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. 2025. [See it from my perspective: How language affects cultural bias in image understanding](#). In *The Thirteenth International Conference on Learning Representations*.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [Vqa: Visual question answering](#). In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Xiaojiao Chen, Yueying Li, Yonghao Chen, Tan Tang, Ruihan Wang, Yifan Wang, Yingchaojie Feng, Wei Chen, and Xiaosong Wang. 2025. [Acpas: an expert-assistance system for authenticating ancient chinese paintings via llm-based agents](#). *npj Heritage Science*, 13(1):512.
- 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. 2025. [Mme: A comprehensive evaluation benchmark for multimodal large language models](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. [Shortcut learning in deep neural networks](#). *Nature Machine Intelligence*, 2(11):665–673.
- Fan Jin, Qingling Chang, and Zhongwen Xu. 2024. [Museumqa: A fine-grained question answering dataset for museums and artifacts](#). In *Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing, MLNLP ’23*, page 221–226, New York, NY, USA. Association for Computing Machinery.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.
- Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, and Mohamed Elhoseiny. 2024. [No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20939–20962, Miami, Florida, USA. Association for Computational Linguistics.
- Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. [Benchmarking vision language models for cultural understanding](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.
- Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025. [Towards cross-lingual explanation of artwork in large-scale vision language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3773–3809, Albuquerque, New Mexico. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Tan Tang, Yanhong Wu, Junming Gao, Kejia Ruan, Yanjie Zhang, Shuainan Ye, Yingcai Wu, and Xiaojiao Chen. 2024. [Arteyer: Enriching gpt-based agents with contextual data visualizations for fine art authentication](#). *Visual Informatics*, 8(4):48–59.
- V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihang Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. [Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning](#). *Preprint*, arXiv:2507.01006.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. [Cider: Consensus-based image description evaluation](#). In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wiki-data: a free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Jing Wan, Xinrong Li, Hao Zhang, Ao Zou, and Rumei Wang. 2025. [An llm-based qa system for chinese painting and calligraphy with knowledge graphs and external documents](#). *npj Heritage Science*, 13(1):643.
- Yuxuan Wang, Yijun Liu, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, Wanxiang Che, and Hongyang Chen. 2025. [Cvluue: a new benchmark dataset for chinese vision-language understanding evaluation](#). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’25/IAAI’25/EAAI’25*. AAAI Press.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). *Preprint*, arXiv:1904.09675.

A Appendix

A.1 CURATORQA Expert Audit and Anti-Shortcut Controls

Expert audit. We audited a stratified random sample of 1,000 CURATORQA questions, stratified by difficulty principle (P1/P2) and QA type (QA1-QA6), using a fixed seed and an artwork-level cap to reduce intra-artwork correlation. An art expert checked gold-answer correctness, ambiguity or multiple-correct-option issues, and principle/type label consistency. The audit found 1 erroneous item (0.1%), which was corrected. Under an exact-binomial calculation, the one-sided 95% upper bound is approximately 0.47%. Because the audit is stratified and uses an artwork-level cap, we treat this value as an audit diagnostic rather than a design-based confidence bound for the full corpus. Table 9 summarizes the audit protocol and result.

Option-bias analysis. Because CURATORQA contains both binary verification items and non-binary multiple-choice items, we report answer-position diagnostics by item format. For binary

Audit item	Value
Audited questions	1,000
Unique artworks represented	785
Sampling scheme	P1/P2 $\times$ QA1-QA6 stratified
Observed erroneous items	1
Observed error rate	0.1%
95% upper bound (diagnostic)	0.47%

Table 9: Expert audit summary for CURATORQA.

verification items ($N = 962$), the A/B distribution is 0.663 / 0.337. For non-binary multiple-choice items ($N = 13,459$), the A/B/C/D distribution is nearly balanced (0.247 / 0.248 / 0.255 / 0.250). Aggregating all 14,421 CURATORQA items gives A/B/C/D = 0.274 / 0.254 / 0.238 / 0.234. Correct options are mildly longer than distractors on average (+0.954 characters), the mean lexical-overlap-rate gap is +0.059, and the template repetition rate is 0.084; Table 10 reports these diagnostics.

Subset	N	A	B	C	D
Binary verification items	962	0.663	0.337	0.000	0.000
Non-binary MCQ items	13,459	0.247	0.248	0.255	0.250
All CURATORQA items	14,421	0.274	0.254	0.238	0.234
Correct-minus-incorrect option length	–			+0.954 chars	
Lexical-overlap-rate gap	–			+0.059	
Template repetition rate	–			0.084	

Table 10: Full-set option-bias diagnostics for CURATORQA. Binary verification and non-binary multiple-choice items are reported separately, and the final row reports the aggregate distribution over all 14,421 questions.

Matched no-image and image-shuffle controls. We run matched controls on representative high-performing models. Full-image accuracy is substantially higher than both no-image and image-shuffled settings, indicating that performance is not explained solely by language priors or image-independent shortcuts (Table 11).

Paired model comparison. Table 12 reports paired and artwork-clustered significance tests for the Qwen3-VL-235B vs. GPT-5-mini comparison discussed in the main text. This table complements the subset-level gap analysis in Table 3.

Model	Full	No img.	Δ	Shuffle	Δ
GPT-5-mini	0.795	0.414	0.381	0.530	0.265
Qwen2.5-VL-72B	0.808	0.489	0.319	0.525	0.283
Qwen3-VL-235B	0.840	0.567	0.273	0.565	0.275
Qwen2.5-VL-32B	0.797	0.236	0.561	0.516	0.279

Table 11: Matched no-image and image-shuffle controls for CURATORQA. Δ denotes the drop from the full-image setting.

A.2 Cross-Institution Protocol Replication

Pilot construction. To examine whether the principal diagnostic patterns of CARTBENCH depend on the Palace Museum source, we constructed a held-out cross-institution protocol-replication pilot using 120 Chinese artworks from five external institutions: the Smithsonian National Museum of Asian Art, the Victoria and Albert Museum, the Art Institute of Chicago, the Cleveland Museum of Art, and the Metropolitan Museum of Art. These institutions differ from the Palace Museum in collection composition, metadata conventions, and curatorial prose.

For External-CURATORQA, we changed the source institutions while preserving the original construction protocol. Official museum descriptions were faithfully translated into Chinese without summarization or factual expansion. We reused the original P1/P2 definitions, QA1-QA6 taxonomy, structural filtering, answer-option cleanup, seeded option shuffling, and exact-match evaluation protocol. The resulting pilot contains 778 questions over 120 artworks, including 565 P1 and 213 P2 items, with all six QA types represented. All 120 translations and all 778 question-answer records received primary human review. We retain 66 open short-answer items separately and evaluate the remaining 712 closed-form items.

External-CURATORQA results. The exact overall model ranking does not reproduce on the external set. GPT-5-mini, Qwen2.5-VL-72B, Qwen3-VL-30B, Qwen3-VL-235B, and Qwen2.5-VL-7B obtain closed-form accuracies of 0.792, 0.789, 0.778, 0.777, and 0.756, respectively. Nevertheless, the principal task-level pattern transfers: P2 is harder than P1 for every tested model, QA5 (style-to-period inference) is the weakest QA type for every model, and the per-model Spearman correlation between the main and external QA-type profiles ranges from 0.81 to 0.99. Four of the five overall accuracies are within 2.2 percentage points of the corresponding main result, whereas Qwen3-VL-

Model A	Model B	Acc. A	Acc. B	Δ	Significance
Qwen3-VL-235B	GPT-5-mini	0.8400	0.7949	+0.0451	McNemar (paired, $N=14,421$): $p=4.22 \times 10^{-47}$. Clustered by artwork ID ($N_{\text{artwork}}=1,589$): $\bar{\Delta}=0.0447$, $CI_{95}=[0.0384, 0.0512]$, $p=0.0002$.

Table 12: Paired comparison between Qwen3-VL-235B and GPT-5-mini on CURATORQA. We report exact-match accuracy, the paired question-level McNemar test, and the artwork ID-clustered analysis. $\Delta = \text{Acc}(A) - \text{Acc}(B)$.

Item	Count	Notes
Artworks	120	Five institutions
Questions	778	All QA1-QA6 covered
P1	565	Visual evidence
P2	213	Vision + art knowledge
Closed-form	712	Exact-match evaluation
Open short-answer	66	Retained separately

Table 13: Summary of the External-CURATORQA pilot. All 120 translations and all 778 question-answer records received primary human review.

235B decreases by 6.3 points. Thus, the pilot supports transfer of the benchmark’s difficulty profile and failure pattern, but not an institution-invariant model ranking.

External-CATALOGCAPTION. We further replicated CATALOGCAPTION on 24 artworks from four external institutions. Art specialists produced and verified a four-section reference for each artwork using the corresponding museum evidence. We applied the original automatic evaluation protocol without changing its metrics, character-level tokenization, or KPI weights.

Qwen3-VL-235B remains above Qwen2.5-VL-7B on the aggregate KPI, whereas Qwen3-VL-235B and GPT-5-mini are effectively tied (0.289 vs. 0.288). The component metrics also yield mixed ordering between these two models. Given the small 24-artwork pilot and the limitations of similarity-based metrics for expert appreciation, we do not interpret these small differences as evidence of a stable cross-institution ranking.

External-REINTERPRET. We also replicated the original two-stage REINTERPRET protocol on 12 artworks from four external institutions. Three art-domain raters blindly evaluated anonymous model outputs using the same Stage 1 plausibility gate and Stage 2 rating protocol as in the main experiment.

On the 11 artworks that majority-pass for both

models, the paired Qwen-minus-GPT Stage 2 difference is +0.300, with an artwork-bootstrap 95% confidence interval of $[+0.091, +0.500]$. For Stage 1, raw agreement is 0.847 and Gwet’s AC1 is 0.799. The three-rater Stage 2 composite has $ICC(2,3) = 0.643$. We therefore emphasize the paired overall direction and uncertainty rather than fine-grained differences between individual rating dimensions.

Scope of the replication. Overall, the cross-institution pilot provides initial evidence that the principal diagnostic patterns identified by CART-BENCH persist when the source institution changes. It does not establish universal generalizability: the exact CURATORQA model ranking changes, the two long-form pilots remain small, and the evaluated objects remain Chinese artworks from a limited set of museum contexts. We therefore interpret these results as evidence of initial cross-institution protocol transfer rather than an invariant model ordering or a universal cross-museum evaluation standard.

A.3 KPI Weight Sensitivity Analysis for CATALOGCAPTION

Motivation. KPI aggregates multiple automatic metrics (BERTScore $_{F1}$, CIDEr-like, ROUGE-L, BLEU-4) into a single score (Equation 1). Since any fixed weighting may appear arbitrary, we test whether our CATALOGCAPTION model ranking is stable under reasonable perturbations of the metric weights.

Protocol. Let $\mathbf{w} = (w_{\text{bert}}, w_{\text{cider}}, w_{\text{rouge}}, w_{\text{bleu}})$ be a nonnegative weight vector with $\sum_i w_i = 1$. For each sampled $\mathbf{w}$, we compute $KPI_{\mathbf{w}} = \sum_i w_i m_i$ where m_i denotes the corresponding metric value, then obtain a model ranking induced by $KPI_{\mathbf{w}}$. We compare each ranking to the default-weight ranking using Spearman’s ρ and Kendall’s τ . We also compute (i) *Top-1 same rate*, the fraction of scenarios in which the best-performing model

Model	KPI $\uparrow$	BERTScore $_{F1}\uparrow$	EmbSim $\uparrow$	CIDEr-like $\uparrow$	ROUGE-L $\uparrow$	BLEU-4 $\uparrow$
Qwen3-VL-235B	0.289	0.367	0.916	0.269	0.223	0.127
GPT-5-mini	0.288	0.360	0.917	0.276	0.226	0.124
Qwen2.5-VL-7B	0.269	0.342	0.896	0.241	0.218	0.113

Table 14: External-CATALOGCAPTION evaluation on 24 artworks from four external institutions. We use the same automatic metrics, tokenization, KPI definition, and metric weights as in the main experiment.

Model	Rating pass	Majority passed	Stage-2 comp. [95% CI]
Qwen3-VL-235B	0.917	11/12	4.285 [4.103, 4.436]
GPT-5-mini	0.889	12/12	3.964 [3.861, 4.064]

Table 15: Primary External-REINTERPRET comparison on 12 artworks with three expert raters. Stage-2 composites are reported with 95% bootstrap confidence intervals.

(excluding human baselines) is unchanged, and (ii) *Top-3 Jaccard*, the Jaccard similarity between the default top-3 model set and the top-3 set under $\mathbf{w}$.

Weight perturbation schemes. We consider the following perturbations: (i) **Uniform**: $w_i = 0.25$ for all metrics. (ii) **Dirichlet**: $\mathbf{w} \sim \text{Dirichlet}(\alpha\mathbf{1})$ with $\alpha \in \{0.5, 1, 5\}$, capturing increasingly concentrated vs. smooth weight distributions. (iii) **Local Gaussian**: sample $\tilde{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}_0, \sigma^2 I)$ around the default weight $\mathbf{w}_0 = (0.45, 0.25, 0.20, 0.10)$ with $\sigma = 0.05$, then project onto the simplex. (iv) **BERT-dominant range**: sample $w_{\text{bert}} \in [0.35, 0.60]$ and distribute the remaining mass across other metrics proportional to the default weights.

Findings. As shown in Table 16, rankings remain highly consistent with the default ordering across a wide range of weight perturbations (median Spearman ≥ 0.964 ; median Kendall ≥ 0.891 even under the more extreme Dirichlet sampling with $\alpha \leq 1$). Table 17 further reports model-level rank stability under Dirichlet perturbations. Crucially, the top-ranked model under the default KPI remains unchanged across all tested scenarios (Top-1 same = 1.0), supporting that our main CATALOGCAPTION conclusions do not hinge on a particular weighting choice.

A.4 Rubric-Based Expert Evaluation for CATALOGCAPTION

Because open-ended appreciation admits multiple valid outputs and automatic similarity metrics may

not fully capture expert insight, we additionally evaluate CATALOGCAPTION with a rubric-based expert study. Three raters score 258 model-artwork pairs (86 artworks $\times$ 3 models) on D1 evidence-grounded interpretation, D2 cultural/art-historical appropriateness, and D3 insight depth using a 1-5 scale, and flag major factual or hallucination errors (E1). The 3-rater mean shows moderate reliability for D1-D3 under ICC(2,3), while the sparse binary E1 flag is better interpreted with Gwet’s AC1 and raw agreement; Table 18 reports the scores and agreement statistics.

A.5 Additional statistics for REINTERPRET

We provide additional breakdowns for REINTERPRET. Table 19 reports dimension-wise macro-averages with bootstrapped 95% confidence intervals, and Table 20 reports Stage 1 gate-failure type distributions.

A.6 Inter-Annotator Agreement for REINTERPRET

We report inter-annotator agreement for both stages of REINTERPRET. For the Stage 1 plausibility gate, we compute Fleiss’ κ over pass/fail judgments. For Stage 2, since the main results use the three-rater mean, we report ICC(2,3) for the averaged Primary score (D1-D3) and All score (D1-D5). The results in Table 21 indicate moderate-to-substantial agreement for the gate and reasonable reliability for averaged Stage 2 scores.

A.7 Human Evaluation Protocol and Annotator Information

Human evaluations were conducted by three domain experts (two art graduate students and one art-school professor in China), all adult native Chinese speakers with formal training in Chinese art history. Annotators were compensated at a reasonable rate meeting or exceeding local minimum wage or typical market rates. Participation was voluntary with research-use notice; we collected no personally identifying information and report only anonymized, aggregate results.

Scenario	#samples	Spearman (p05/p50/p95)	Kendall (p05/p50/p95)	Top-1 same	Top-3 Jaccard (p05/p50/p95)
Default (sanity)	1	1.000/1.000/1.000	1.000/1.000/1.000	1.000	1.000/1.000/1.000
Uniform weights	1	0.991/0.991/0.991	0.964/0.964/0.964	1.000	1.000/1.000/1.000
Dirichlet $\alpha=5$	8000	0.945/0.991/0.991	0.855/0.964/0.964	1.000	1.000/1.000/1.000
Dirichlet $\alpha=1$	8000	0.882/0.964/1.000	0.709/0.891/1.000	1.000	0.500/1.000/1.000
Dirichlet $\alpha=0.5$	8000	0.873/0.964/1.000	0.673/0.891/1.000	1.000	0.500/1.000/1.000
$w_{\text{bert}} \in [0.35, 0.60]$	5000	0.991/1.000/1.000	0.964/1.000/1.000	1.000	1.000/1.000/1.000
Local Gaussian $\sigma=0.05$	5000	0.982/0.991/1.000	0.927/0.964/1.000	1.000	0.500/1.000/1.000

Table 16: KPI weight sensitivity analysis for CATALOGCAPTION. Rank correlations are computed against the default-weight ordering; we report 5th/50th/95th percentiles for stochastic scenarios. Top-1 same and Top-3 Jaccard are computed over *models* (excluding human baselines).

Model	Mean rank	Std. rank	$P(\text{top-3})$
human_mean	1.000	0.000	1.000
preds_qwen25_72B	2.397	0.734	0.860
preds_qwen3_235B	2.863	0.495	0.936
preds_qwen3_30B	4.035	0.806	0.203
preds_qwen25_7B	5.038	0.920	0.000
preds_glm45	6.189	0.785	0.000
preds_glm46	6.764	0.766	0.000
preds_qwen25_32B	8.373	1.083	0.000
gpt5nano	8.733	0.860	0.000
gpt5mini	9.608	0.509	0.000
gemini2.5	11.000	0.000	0.000

Table 17: Model rank stability under Dirichlet weight perturbations with $\alpha=1$ (8,000 samples). We report the mean and standard deviation of each model’s rank, and the probability of appearing in the top-3.

Item / Model	D1	D2	D3	E1
Inter-rater agreement	ICC=0.639	ICC=0.622	ICC=0.637	$\kappa=0.012$; AC1=0.841; Agr.=0.863
Qwen2.5-VL-72B	3.640	3.702	3.663	0.081
Qwen3-VL-235B	3.647	3.609	3.616	0.050
Gemini-2.5-Flash	3.229	3.275	3.337	0.093

Table 18: Rubric-based expert evaluation for CATALOGCAPTION. D1-D3 are 1-5 scores; E1 is the major factual/hallucination error rate.

A.8 Preliminary expert error attribution for CONNOISSEURPAIRS

We summarize a preliminary expert attribution on misclassified cases in CONNOISSEURPAIRS. Errors are consistently associated with weak sensitivity to period/style consistency, overweighting local cues such as inscriptions and seals, and relying on surface-level detail richness rather than global brush-and-ink coherence.

A.9 Reproducibility Details

We will release metadata, task prompts, derived annotations, model outputs used for scoring, and

evaluation scripts for exact match, automatic caption metrics, bootstrap confidence intervals, Wilson intervals, and human-evaluation aggregation. We do not redistribute Palace Museum images or curatorial texts; instead, the release provides stable identifiers and source URLs so users can access media subject to the source institutions’ terms. The release also includes validation scripts for missing URLs, duplicated IDs, answer-option balance, and split consistency.

A.10 Prompt for CURATORQA

Chinese prompt.

你是中国艺术VQA 基准数据集的题目设计员。你只拿到“描述文本”，但你要设计在评测时由VLM 通过图像（P1）或图像+常识（P2）回答的题目。

(You are a question designer for a Chinese-art VQA benchmark. You only have access to a textual description, but you must design questions answerable by a VLM using the image (P1) or image + commonsense (P2).)

硬性约束:

(Hard constraints:)

- 只输出P1 或P2。(Output only P1 or P2.)
- 不要生成需要依赖“标题/作者/年代/著录/历史事件细节/释文”等文本信息才能回答的题。(Do not generate questions requiring title/author/date/provenance/historical details/transcription.)
- 每题必须给出2到5条预期可见线索，并提供一句说明为什么不需要阅读描述文本也能回答。(Each question must provide 2-5 expected visual cues and a one-sentence explanation of why it is answerable without the text.)
- 只能写“预期的图像证据/可见线索”，不得引用或复述描述原句。(Only write expected image evidence/visual cues; do not quote the description.)

题型覆盖要求：生成6到10题，只允许P1或P2，并尽量覆盖：

(Coverage requirement: Generate 6-10 questions, P1/P2 only, covering:)

Model	Gate pass	D1 Novelty	D2 Coherence	D3 Evidence	D4 Elaboration	D5 Insight
gemini-2.5-flash	0.707 0.813]	[0.600, 3.287 [2.940, 3.653]	3.960 [3.740, 4.193]	3.773 [3.587, 3.960]	3.980 [3.760, 4.213]	3.613 [3.353, 3.907]
GLM-4.6V	0.853 0.933]	[0.760, 3.407 [3.213, 3.620]	3.793 [3.653, 3.947]	3.747 [3.560, 3.947]	3.800 [3.627, 3.980]	3.647 [3.467, 3.853]
gpt-5-mini	0.867 0.947]	[0.787, 3.280 [3.120, 3.447]	3.767 [3.647, 3.880]	3.653 [3.507, 3.787]	3.873 [3.733, 4.027]	3.513 [3.333, 3.700]
qwen3-vl-235B	0.880 0.960]	[0.787, 3.533 [3.340, 3.740]	3.867 [3.687, 4.053]	3.840 [3.707, 3.973]	3.847 [3.693, 4.013]	3.553 [3.413, 3.700]

Table 19: Dimension-wise macro-averages for REINTERPRET with 95% bootstrap confidence intervals over artworks. Gate pass is the Stage 1 plausibility pass rate. D1-D5 are Stage 2 Likert ratings computed only for gate-passed outputs, macro-averaged over artworks.

Model	Failures	Other	Era/Dynasty error	Misread / missed content	Wrong type/region attribution	Empty/NA
gemini-2.5-flash	22 (29.3%)	11 (14.7%)	6 (8.0%)	2 (2.7%)	2 (2.7%)	1 (1.3%)
GLM-4.6V	11 (14.7%)	8 (10.7%)	2 (2.7%)	1 (1.3%)	0 (0.0%)	0 (0.0%)
gpt-5-mini	10 (13.3%)	5 (6.7%)	1 (1.3%)	3 (4.0%)	1 (1.3%)	0 (0.0%)
qwen3-vl-235B	9 (12.0%)	7 (9.3%)	1 (1.3%)	0 (0.0%)	1 (1.3%)	0 (0.0%)

Table 20: Stage 1 gate-failure type distributions for REINTERPRET. Counts are over rater×artwork instances (75 per model), with percentages relative to 75. Failure types are recorded only when a rater marks the output as failing the plausibility gate.

Model	Gate κ	ICC Primary	ICC All
GLM-4.6V	0.487	0.498	0.649
gemini-2.5-flash	0.415	0.476	0.709
gpt-5-mini	0.711	0.785	0.869
qwen3-vl-235B	0.577	0.877	0.888

Table 21: Inter-annotator agreement for REINTERPRET. Stage 1 uses Fleiss’ κ ; Stage 2 reports ICC(2,3) for three-rater averaged scores.

- QA1 主体/对象识别 (Subject/object recognition, P1)
- QA2 场景/活动归类 (Scene/activity classification, P2)
- QA3 形制/形式 (Composition/format: scroll/album/fan, P1/P2)
- QA4 技法线索 (Visible technique/style cues, P1)
- QA5 风格→时期推断 (Style→period inference; only when description implies visible cues; must be P2; phrase as “which period is more likely” with “unable to determine” option)
- QA6 可见符号/对象 (Visible iconography/elements, P1)

A.11 Prompt for CATALOGCAPTION

Chinese prompt.

你将收到一张中国艺术作品的图片。请用中文输出“专业鉴赏式”描述，并严格遵循以下格式与要求：

(You will be given an image of a Chinese artwork. Write a professional, connoisseur-style appreciation in Chinese, strictly following the format below.)

格式要求：必须按顺序输出四个小节标题，且每节1-3句：

(**Format:** Output four section headers in order, 1-3 sentences each:)

- 【作品背景】 ([Artwork Background]: Make reasonable inferences; if uncertain, state “hard to determine from the image alone.” Do not fabricate author/site/institution.)
- 【作品内容】 ([Artwork Content]: Describe subject, theme, composition, key visible details.)
- 【艺术特色】 ([Artistic Characteristics]: Discuss material, technique, color, lines, style cues from visible evidence.)
- 【综合评价】 ([Overall Evaluation]: Provide aesthetic judgment explaining why it is compelling. Avoid generic praise.)

约束：若某信息无法从图像可靠判断，请明确说明不确定，不要编造。语言要求专业、具体、有依据。总长度建议不超过1200字。

(**Constraints:** If uncertain, state so explicitly. Keep writing professional, specific, evidence-grounded. Recommended length: ≤ 1200 characters.)

A.12 Prompt for REINTERPRET

Chinese prompt.

你将收到一张艺术品图片。请用中文写一段“创新赏析”风格的鉴赏文本，尽量依据可见内容，不要编造无法从图像判断的具体作者、出土地点或藏馆信息；不确定就说明“仅凭图像难以确定”。

(You will be given an artwork image. Write an “innovative appreciation” in Chinese, grounded in visible evidence. Do not invent specific facts not determinable from the image; if uncertain, state so.)

输出请严格包含以下四个小节标题（每节1至3句，信息密度高）：

(Output must include four section headers, 1-3 sentences each, high information density:)

- 【背景】 ([Background]: era, use, cultural context; reasonable inference allowed)
- 【作品内容】 ([Content]: main subject, theme, composition, key visual details)
- 【艺术特色】 ([Artistic Characteristics]: material, craft, technique, linework, coloration, style)
- 【意义价值】 ([Significance]: artistic value, historical/cultural significance, influence)

总长度建议不超过<MAX_CHARS> 字。

(Recommended length: $\leq$ <MAX_CHARS> characters.)

A.13 Prompt for CONNOISSEURPAIRS

Chinese prompt.

你将看到一张对比图：左侧为图像A，右侧为图像B（图上角标已标注A/B）。两幅图为同一题材作品的一件原作/真本与一件摹本、仿本或复制相关作品。

(You will see a side-by-side comparison: Image A on the left, Image B on the right. The pair contains one original work and one copy-, imitation-, or reproduction-related work of a similar subject.)

任务：判断【哪一张更可能是原作/真本】（只能选A或B）。

(**Task:** Decide which side is more likely to be the original. Choose only A or B.)

依据限制：只能依据图像证据作判断，不得依赖标题、作者、年代、收藏信息等外部知识。

(**Restriction:** Base decision only on visible evidence. Do not use external metadata.)

输出要求：

(**Output:**)

- (1) 结论 (Decision): A or B.
- (2) 理由 (Reasons): 3-6 points grounded in visual cues (brushwork, ink layering, texture strokes, color, support texture, seal consistency, copying artifacts).
- (3) 不确定性 (Uncertainty): If evidence is close, state the key uncertainty and what additional visual details would be needed.

A.14 Survey instrument for REINTERPRET

We evaluate REINTERPRET with a two-stage human questionnaire designed for Chinese-art “creative reinterpretation.” Figures 3-4 show the original Chinese questionnaire. Below we provide an item-by-item English translation of the questionnaire text. Raters are presented with an artwork image and an anonymous model-generated interpretation.

Literal English translation.

1. **Title and opening note.** Model creativity human-evaluation form. Welcome to this large-scale human evaluation. You will sequentially evaluate images of 25 canonical works from Chinese art history and the corresponding anonymous appreciation texts. Each work is evaluated independently in two stages: first judging the text’s basic reasonableness and eligibility to enter the creativity evaluation, and then, for texts that pass the screening, giving detailed scores on creativity-related dimensions. Please complete each item according to the specific content of each work and the instructions. Your professional judgment will provide an important reference for evaluating the creativity of model-generated content.
2. **Evaluation information.** Rater ID: _____.
3. **Instructions (please read first).** Welcome to this human evaluation. You will evaluate an image of a canonical work from Chinese art history and an anonymous appreciation text corresponding to it. This evaluation adopts a two-stage procedure: Stage 1 is a basic correctness/reasonableness screening; Stage 2 is a creativity evaluation, scored only for texts that pass Stage 1. Please note: this evaluation does not require the text to repeat authoritative conclusions, but the text should not obviously violate basic art-historical consensus. Creativity should be built on observable image information and reasonable interpretation. _____ (please fill in “read”).
4. **Artwork 1: basic reasonableness judgment [single choice].** Does the text contain any of the following problems: obvious misreading of the main content/theme of the image; clearly wrong judgment of the work’s type/period/attributes; obvious violation of art-historical consensus; or large amounts of fabrication without marking uncertainty, such as a specific author or date? Options: *yes* (the above problem exists; do not enter the second stage) / *no* (no problem found; enter the second stage).
5. **Artwork 1: if selecting “yes,” please explain the problem type (optional) [multi-field fill-in].** Main problem type: _____.
6. **Artwork 1: Interpretive Novelty [single choice].** Does it go beyond a template-like appreciation style and propose a new idea in

interpretive angle or meaning association? Options: 1 = basically textbook-style statement; 2 = mainly follows a conventional path, with only surface-level variation; 3 = some content reflects novelty; 4 = proposes a clear new interpretive angle; 5 = forms a significantly different interpretive reorganization.

7. **Artwork 1: Integrative Coherence [single choice].** Do the background, content, features, and value sections form a logically coherent interpretive structure? Options: 1 = the parts are clearly disconnected; 2 = the connection is weak and overall integrity is insufficient; 3 = basically coherent but with limited structural function; 4 = the structure is clear and the contents support one another; 5 = forms a complete interpretive chain.
8. **Artwork 1: Evidence-based Reasoning [single choice].** Judge whether the interpretation is mainly based on observable image information and reasonably handles uncertainty. Options: 1 = many judgments without evidence; 2 = there are inappropriate inferences without limiting uncertainty; 3 = the inferences are reasonable but the evidence is general; 4 = the inferences are clear and the evidence is explicit; 5 = the inferences are rigorous and the relationship between evidence and conclusion is clear.
9. **Artwork 1: Elaboration & Expressiveness [single choice].** Is the expression concrete and layered, avoiding empty evaluation? Options: 1 = expression is empty; 2 = relatively general, lacking detail; 3 = has some specificity; 4 = expression is concrete and clearly layered; 5 = expression is concise and information-dense.
10. **Artwork 1: Perceived Creative Insight [single choice].** Does it provide new thinking clues or discussion space for understanding the work? Options: 1 = no inspiration; 2 = weak inspiration; 3 = has some inspirational value; 4 = provides a clear direction for extension; 5 = has significant inspirational potential.
11. **Artwork 1: rater supplementary comments (optional).** Particularly outstanding creativity, typical problems, or interpretations worth discussing: _____.
12. **Artwork 1: rater reminder.** Note: the goal of this evaluation is not to judge whether the text “correctly repeats authoritative conclusions,” but to evaluate whether, under the premise that basic

understanding is established, the text conducts grounded creative reinterpretation.

A.15 Expert error-analysis questionnaire for CONNOISSEURPAIRS

For CONNOISSEURPAIRS, we conduct an expert-facing error analysis on misclassified original-copy pairs. Figures 5-6 show the original Chinese questionnaire. Below we provide an item-by-item English translation. In this appendix, terms such as “authentic” appear only as part of the annotator-facing survey wording; throughout the benchmark, the task is treated as original-copy discrimination rather than a real-world authentication protocol.

Literal English translation.

1. **Title and opening note.** Questionnaire for analyzing model error cases in artwork authenticity/connoisseurship. Welcome to this questionnaire. This questionnaire aims to collect your professional insights into model error cases in the artwork authenticity/connoisseurship task. Please analyze the reasons for the model’s error from a professional or intuitive perspective according to the given artwork pair and model decision result. Your feedback is very important and will be used for error classification, error abstraction-level analysis, and related research content in the paper. Some questions are multiple choice, some are single choice, and there is also one open-ended question for briefly summarizing the model problem. Thank you for your support and cooperation.
2. **Model 1-Artwork 1 decision.**
3. **Model decision [single choice].** The model’s judgment result; gold indicates the real/authentic work, and pred indicates the work predicted by the model as real/authentic. Options: A is judged by the model as real/authentic; B is judged by the model as real/authentic.
4. **Main reasons for the model error [multiple choice].** Which cognitive or visual errors do you think the model most likely made? Options: over-focused on local details such as brushstroke density, line complexity, and texture richness; ignored global composition and style consistency, where local parts appear reasonable but the overall spirit/resonance and layout do not hold; mistook “finely painted / complex” for “older / more authentic”; was misled by a

high-quality copy or reproduction, where the reproduction is technically close to the original; could not understand the brush-and-ink logic of Chinese painting, such as central-tip / side-tip brushwork, brush rhythm, and void-solid relations; gave excessive weight to inscriptions, seals, or textual elements; was insensitive to period-style differences, being unable to distinguish the aesthetics and techniques of different periods; made an overall judgment more like surface-similarity matching than art-historical principles; other.

5. **Other error-cause explanation.** If selecting “other,” please briefly explain the reason: _____.
6. **Model error layer [single choice].** From the root, which level of ability is insufficient? Options: low-level visual problem, meaning failure to correctly use visual cues such as lines, structure, and space; mid-level style-modeling problem, meaning lack of modeling of “style,” “spirit/resonance,” and “overall consistency”; high-level cultural/art-historical understanding problem, meaning lack of understanding of the logic and historical context of Chinese painting connoisseurship; multiple layers are involved and it is difficult to attribute the error to a single level.
7. **Error keyword summary [single choice].** If you could choose only one keyword to summarize the error, which would you choose? Options: local-detail bias, meaning over-reliance on local visual features such as brushstrokes, line density, and texture; global-style insensitivity, meaning failure to recognize consistency in composition, spirit/resonance, and style; copy mistaken as authentic, meaning being misled by a high-quality copy or reproduction; inscription/seal over-reliance, meaning giving excessive weight to textual or seal information; period-style insensitivity, meaning inability to distinguish aesthetic and technical features of different periods; surface-similarity shortcut judgment, meaning judging based on overall visual similarity rather than art-historical principles; other.
8. **Human avoidability [single choice].** Is this error “easy for humans to avoid”? Options: yes, humans, especially experienced people, can avoid it easily; average, humans also need sub-

stantial experience; no, this case is also challenging for humans.

9. **Model problem summary.** Please summarize the model’s core problem in this sample in one sentence, no more than 30 Chinese characters: _____.

A.16 Eligibility Audit of Chinese-Specialized Baselines

We screened publicly documented candidate systems before inference using four task-derived requirements: native image input, generative output, support for all four CARTBENCH protocols, and sufficient generality across painting, calligraphy, sculpture, ceramics, and stationery. The audited candidates fell into three recurring categories: image-conditioned generative systems that are bilingual or multilingual rather than strictly Chinese-monolingual; non-generative image-text encoders; and heritage-specific systems whose scope does not cover all four protocols or represented art forms. We therefore do not treat any of these systems as a directly comparable full-benchmark monolingual Chinese baseline.

A.17 LLM Configuration

Access methods. OpenAI models were accessed via the OpenAI API.⁵ Gemini models were accessed via Google’s Gemini API.⁶ For the open-weight Qwen and GLM families, we used the AI Ping API gateway (<https://www.aiping.cn/>) and selected model identifiers from its public model list endpoint.⁷

Model identifiers used in our experiments. OpenAI models include gpt-5-mini and gpt-5-nano for evaluation, and gpt-5.2 for LLM-assisted CURATORQA question generation. Gemini uses gemini-2.5-flash. AI Ping model IDs include qwen3-v1-235b-a22b-thinking, qwen3-v1-30b, qwen2.5-v1-72b, qwen2.5-v1-32b, qwen2.5-v1-7b, glm-4.5v, and glm-4.6v.

Decoding and inference hyperparameters. We request temperature=0 and top_p=1 where supported and generate a single response per prompt. These settings are intended to reduce sampling variability but do not guarantee exact determinism for

⁵<https://platform.openai.com/docs/overview>

⁶<https://ai.google.dev/gemini-api/docs>

⁷<https://www.aiping.cn/docs/API/modelList>

models accessed through hosted APIs or gateways. Unless otherwise noted, the configured maximum output length is 16k tokens.

A.18 Data stats

Table 22 reports the main statistics and distributions.

A.19 Dynasty timeline used for era statistics

To make the era distribution in Figure 2 interpretable to readers unfamiliar with Chinese history, we provide in Figure 7 a reference timeline of the dynasties and historical periods used in this work. We follow a widely adopted periodization in Chinese art history and use these ranges as *coarse* bins when reporting era coverage. Here, *BCE* denotes “Before Common Era” and *CE* denotes “Common Era”; *c.* (circa) indicates an approximate date range. Some periods correspond to transitional eras or partially overlapping regimes (e.g., the Sixteen Kingdoms period), and museum records may use finer-grained attributions. For consistency in our analysis, we map such metadata to the closest period in Figure 7.

Group	Tag	Description	Numbers	Percentage
Overview				
Entries	–	Unique artwork entries	1,589	–
Questions	–	Total number of questions	14,421	–
Avg. Q / entry	–	Average questions per entry	9.07	–
Difficulty				
Principle	P1	Visual-only (answerable from the image)	10,920	75.72
Principle	P2	Vision plus art commonsense/knowledge	3,501	24.27
Question types				
QA type	QA1	Subject recognition (primary subject/object)	2,135	14.80
QA type	QA2	Scene/activity classification	1,932	13.39
QA type	QA3	Composition/format (form, layout, format)	2,578	17.87
QA type	QA4	Visible technique/style cues	2,882	19.98
QA type	QA5	Style→period inference	1,493	10.35
QA type	QA6	Iconography (visible symbols/elements)	3,401	23.58

Table 22: Summary statistics of CURATORQA, including overview counts, difficulty principles, and question-type distribution.

模型创造力人工评测表

欢迎参与本次大规模人工评测！您将依次对 25 幅中国美术史经典作品的图像及对应的匿名鉴赏文本进行评估。每幅作品将独立进行两阶段评测：首先判断文本的基本合理性与可进入创造力评测的资格，随后对通过筛查的文本在创造力相关维度进行详细评分。请根据每幅作品的具体内容，按照指引逐项完成评测。您的专业判断将为模型生成内容的创造性评估提供重要参考。

* 01 评测信息

评审编号：_____

* 02 评测说明（请先阅读）

欢迎参与本次人工评测。您将对一幅中国美术史经典作品的图像及一段对应的匿名鉴赏文本进行评估。本评测采用两阶段评测方式：阶段一：基本正确性/合理性筛查；阶段二：创造力评测（仅对通过阶段一的文本评分）。请注意：本评测不要求文本复述权威结论，但不应明显违背艺术史基本共识。创造性应建立在可观察图像信息与合理解释之上。_____（请填写‘已阅读’）

* 03 艺术品 1：基本合理性判断[单选题]

该文本是否存在以下任一问题：明显误读画面主要内容/题材；明显错误的作品类型/时代/属性判断；明显违背艺术史共识；大量未标注不确定性的编造\（如具体作者、年代等\）

☐ 是（存在上述问题，不进入第二阶段）

☐ 否（未发现问题，进入第二阶段）

* 04 艺术品 1：如选择‘是’，请说明问题类型（可选）[多项填空题]

主要问题类型：_____

* 05 艺术品 1：解释创新性（Interpretive Novelty）[单选题]

是否超越模板化鉴赏写法，在解释角度或意义关联上提出新思路？

☐ 01 基本为教科书式表述

☐ 02 常规路径为主，仅有表层变化

☐ 03 部分内容体现新意

☐ 04 提出清晰的新解释角度

☐ 05 形成显著不同的解释重组

* 06 艺术品 1：解释整合度（Integrative Coherence）[单选题]

背景/内容/特色/价值四部分是否形成逻辑连贯的解释结构？

☐ 01 各部分明显割裂

☐ 02 关联较弱，整体性不足

☐ 03 基本连贯但结构作用有限

☐ 04 结构清晰，内容相互支撑

☐ 05 形成完整解释链条

* 07 艺术品 1：基于图像的推断质量（Evidence-based Reasoning）[单选题]

判断是否主要基于可观察图像信息并合理处理不确定性？

☐ 01 多处无依据判断

☐ 02 存在不当推断未限定不确定性

☐ 03 推断合理但依据概括

☐ 04 推断清楚，依据明确

☐ 05 推断严谨，依据结论关系清晰

Figure 3: Survey questionnaire for REINTERPRET-Part 1, covering evaluation instructions, the Stage 1 plausibility gate, issue-type notes, and the first three creativity dimensions.

*** 08 艺术品 1：表达精致性 (Elaboration & Expressiveness) [单选题]**

是否具体有层次，避免空泛评价？

- ☐ 1 表达空泛
- ☐ 2 偏概括，细节不足
- ☐ 3 有一定具体性
- ☐ 4 表达具体、层次清楚
- ☐ 5 表达精炼且信息密度高

*** 09 艺术品 1：整体创新赏析启发性 (Perceived Creative Insight) [单选题]**

是否为理解作品提供新的思考线索或讨论空间？

- ☐ 1 无启发
- ☐ 2 启发性较弱
- ☐ 3 具有一定启发价值
- ☐ 4 提供清晰延展方向
- ☐ 5 具有显著启发潜力

*** 10 艺术品 1：评审补充意见 (可选)**

特别突出的创造性/典型问题/值得讨论的解读： _____

*** 11 艺术品 1：评审提示**

注意：本评测目标不是判断'正确复述权威结论'，而是评估在基本理解成立前提下，是否进行有根据的创造性再解释。

Figure 4: Survey questionnaire for REINTERPRET-Part 2, covering D4 elaboration and expressiveness, D5 perceived creative insight, optional comments on notable strengths or issues, and a reminder that creativity is evaluated only after the basic plausibility gate is passed.

艺术作品真假品鉴定模型错误案例分析问卷

欢迎参与本次问卷！本问卷旨在收集您对艺术作品真假品鉴定任务中模型错误案例的专业见解。请您根据给定的艺术作品对及模型判断结果，从专业或直观角度分析模型出错的原因。您的反馈对我们非常重要，将用于论文中的错误分类、错误抽象层次分析等研究内容。问卷中部分问题为多选，部分为单选，还有一道开放题供您简要总结模型问题。感谢您的支持与配合！

* 01 模型 1-艺术品 1 的决策

* 02 模型的决策[单选题]

模型的判断结果， gold 表示真的作品， pred 表示模型预测的作品为真

☐ A 被模型判为真

☐ B 被模型判为真

* 03 模型错误的主要原因[多选题]

你认为模型最可能犯了哪些认知或视觉错误？（多选）

☐ 过度关注局部细节（如笔触密度、线条复杂度、纹理丰富程度）

☐ 忽略整体构图与风格一致性（局部看似合理，但整体气韵、布局不成立）

☐ 将“画得精细/复杂”误判为“更古/更真”

☐ 被高质量临摹或复制品欺骗（复制品在技术层面接近真迹）

☐ 无法理解中国书画的笔墨逻辑（如中锋/侧锋、用笔节奏、虚实关系）

☐ 对题跋、印章或文字元素赋予了过高权重

☐ 对时代风格差异不敏感（无法区分不同时期的审美与技法特征）

☐ 整体判断更像基于“表面相似度”而非艺术规律

☐ 其他

* 04 其他错误原因说明

若选择'其他'，请简要说明原因

* 05 模型错误层级[单选题]

从根本上看，这是哪一层面的能力不足？（单选）

☐ 低层视觉问题（未能正确利用线条、结构、空间等视觉线索）

☐ 中层风格建模问题（缺乏对“风格”“气韵”“整体一致性”的建模）

☐ 高层文化/艺术史理解问题（不了解中国书画的鉴定逻辑与历史语境）

☐ 多层问题叠加，难以归因于单一层级

* 06 错误关键词概括[单选题]

如果只能选一个关键词概括该错误，你会选？（单选）

☐ 局部细节偏置（过度依赖笔触、线条密度、纹理等局部视觉特征）

☐ 整体风格不敏感（未能识别构图、气韵与风格的一致性）

☐ 临摹误判为真迹（被高质量临摹或复制品所误导）

☐ 题跋 / 印章过度依赖（对文字或印鉴信息赋予过高权重）

☐ 时代风格不敏感（无法区分不同时代的审美与技法特征）

☐ 表面相似性捷径判断（基于整体外观相似度而非艺术规律作出判断）

☐ 其他

Figure 5: Survey questionnaire for CONNOISSEURPAIRS-Part 1, covering model-decision verification, multi-select attribution of likely error causes, the primary abstraction layer of the error, and a one-keyword summary of the failure mode.

*** 07 人类避免难度[单选题]**

该错误是否“人类容易避免”？（单选）

☐ 是，人类（尤其有经验者）很容易避免

☐ 一般，人类也需要较多经验

☐ 否，这类情况对人类也具有挑战性

*** 08 模型问题总结**

请用一句话概括模型在该样本中的核心问题（不超过 30 字）

Figure 6: Survey questionnaire for CONNOISSEURPAIRS-Part 2, covering whether the error would be easy for human experts to avoid and a one-sentence diagnostic summary of the model's core failure.

中国朝代/Corresponding English	时间/time
新石器时代/Neolithic period, China	c. 10,000 - 2,000 BCE
商/Shang Dynasty	c. 1,600 - 1,046 BCE
西周/Western Zhou Dynasty	1,046 - 771 BCE
战国/Warring States period	c. 475 - 221 BCE
秦/Qin Dynasty	221 - 206 BCE
汉/Han Dynasty	202 BCE - 220 CE
三国/Three Kingdoms period	220 - 280 CE
晋/Jin Dynasty	265 - 420 CE
十六国/Sixteen Kingdoms period	304 - 439 CE
南北朝/Northern and Southern Dynasties	420 - 589 CE
隋/Sui Dynasty	581 - 618 CE
唐/Tang Dynasty	618 - 907 CE
五代十国/Five Dynasties and Ten Kingdoms period	907 - 960 CE
辽/Liao Dynasty	916 - 1125 CE
宋/Song Dynasty	960 - 1279 CE
西夏/Western Xia Dynasty	1038 - 1227 CE
金/Jin Dynasty	1115 - 1234 CE
元/Yuan Dynasty	1271 - 1368 CE
明/Ming Dynasty	1368 - 1644 CE
清/Qing Dynasty	1644 - 1912 CE
近现代/Modern & Contemporary China	1912 CE - present

Figure 7: Chinese dynasties and historical periods considered in this work, listing the Chinese names with English equivalents and their approximate date ranges (BCE/CE).