CiQi-Agent: Aligning Vision, Tools and Aesthetics in Multimodal Agent for Cultural Reasoning on Chinese Porcelains

Wenhan Wang[1,2*], Zhixiang Zhou[2*], Zhongtian Ma[3*], Yanzhu Chen[2], Ziyu Lin[2], Hao Sheng[1], Pengfei Liu[2], Wenqi Shao[2†], Qiaosheng Zhang[2,3†], and Yu Qiao[2,3]

[1] Beihang University
[2] Shanghai Innovation Institute
[3] Shanghai Artificial Intelligence Laboratory
{wqshao,zhangqiaosheng}@sii.edu.cn
https://github.com/SII-Monument-Valley/CiQi-Agent

Abstract. The connoisseurship of antique Chinese porcelain demands extensive historical expertise, material understanding, and aesthetic sensitivity, making it difficult for non-specialists to engage. To democratize cultural-heritage understanding and assist expert connoisseurship, we introduce **CiQi-Agent**—a domain-specific **Porcelain Connoisseurship Agent** for intelligent analysis of antique Chinese porcelain. CiQi-Agent supports multi-image porcelain inputs and enables vision tool invocation and multimodal retrieval-augmented generation, performing fine-grained connoisseurship analysis across **six attributes**: dynasty, reign period, kiln site, glaze color, decorative motif, and vessel shape. Beyond attribute classification, it captures subtle visual details, retrieves relevant domain knowledge, and integrates visual and textual evidence to produce coherent, explainable connoisseurship descriptions. To achieve this capability, we construct a large-scale, expert-annotated dataset **CiQi-VQA**, comprising 29,596 porcelain specimens, 51,553 images, and 557,940 visual question–answering pairs, and further establish a comprehensive benchmark **CiQi-Bench** aligned with the previously mentioned six attributes. CiQi-Agent is trained through supervised fine-tuning, reinforcement learning, and a tool-augmented reasoning framework that integrates two categories of tools: a vision tool and multimodal retrieval tools. Experimental results show that CiQi-Agent (7B) outperforms all competitive open- and closed-source models across all six attributes on CiQi-Bench, achieving on average **12.2%** higher accuracy than GPT-5.

Keywords: LLM-based Agent · Multimodal Dataset & Benchmark · Chinese Porcelain

[*] Equal contribution.
[†] Corresponding authors.

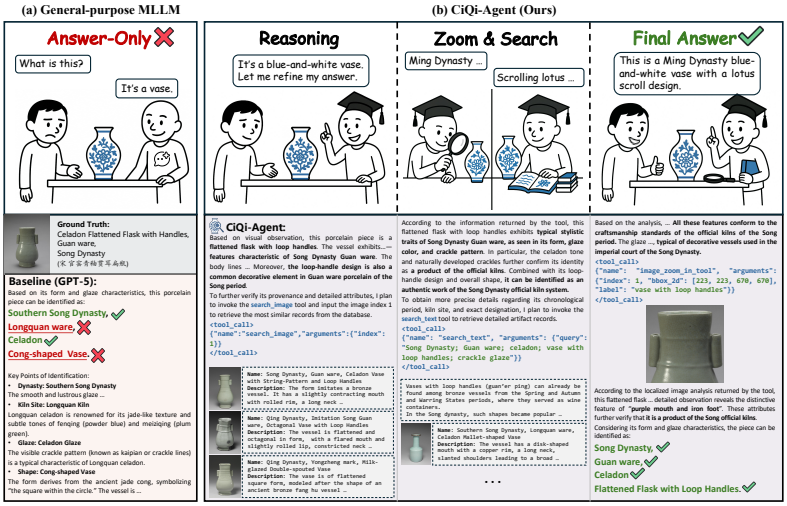

Fig. 1: Comparison between (a) General-purpose MLLM and (b) CiQi-Agent. (a) Conventional MLLMs rely on *single-pass answering*, directly outputting a label leading to superficial or inaccurate identifications. (b) The proposed **CiQi-Agent** introduces *Tool-Augmented Reasoning*, enabling multi-step porcelain analysis through image zoom-in, image/text retrieval. This iterative process yields more reliable final answers aligned with expert connoisseurship reasoning.

1 Introduction

Cultural relics are invaluable carriers of human civilization, encapsulating both artistic creation and historical evolution. The connoisseurship and authentication of artifacts require deep expertise—combining historical knowledge, perceptual experience, and material understanding. Consequently, professional barriers have long restricted public engagement in cultural heritage connoisseurship. With the rapid progress of artificial intelligence (AI) methods, especially large language models (LLMs) and multimodal large language models (MLLMs) [4,14, 16, 31], new opportunities have emerged to democratize artifact understanding. By jointly modeling vision, language, and reasoning, MLLMs enable interactive, explainable, and scalable analysis of visual art and historical objects [2,11,20,36]. In this work, we focus on **Antique Chinese Porcelain**, one of the most representative yet technically challenging categories of cultural relics, as our entry point for AI-driven artifact connoisseurship.

Existing approaches to porcelain connoisseurship remain limited in several key aspects. In most computer vision (CV) studies, porcelain connoisseurship is simplified as a fine-grained recognition task [10, 15, 17, 18], focusing mainly on visual classification without incorporating language-based reasoning or interactive explanation. Moreover, general-purpose MLLMs perform poorly in this specialized domain due to the scarcity of porcelain-related data, leading to weak generalization, unstable judgments, and the absence of a unified evaluation standard,

resulting in outputs that are often incomplete and lacking in professionalism. These limitations highlight the need for a dedicated *porcelain connoisseurship agent* that integrates perception, reasoning, and cultural interpretation.

Building such an agent presents several unique challenges. First, the number of valuable porcelain artifacts is inherently limited, making high-quality data collection extremely difficult. Second, accurate annotation and description of porcelain pieces require deep domain expertise in art history and craftsmanship, which makes large-scale labeling both costly and inconsistent. Third, there is no unified evaluation standard for porcelain connoisseurship, leaving MLLMs without reliable metrics to assess their performance. Finally, the essence of porcelain connoisseurship lies in *fine-grained recognition*, which demands precise identification of features such as vessel shape, glaze color, decorative motifs, and historical period—a task that remains highly challenging even for human experts.

To overcome these challenges, we propose **CiQi-Agent**, the first Chinese porcelain connoisseurship agent that integrates domain-grounded data curation, a two-phase training paradigm combining supervised fine-tuning (SFT) and reinforcement learning (RL), and tool-augmented reasoning into a unified framework (as shown in Fig. 1). CiQi-Agent supports multi-image input and performs fine-grained connoisseurship over six attributes (dynasty, reign period, kiln site, glaze color, decorative motif and vessel shape) by natively integrating vision tool invocation and multimodal retrieval-augmented generation (RAG) as essential reasoning capabilities, and can optionally output the top-k visually similar porcelains to facilitate human interpretation and reference. Specifically, our contributions are as follows:

- We construct **CiQi-VQA**, a large-scale dataset for ancient Chinese porcelain connoisseurship, with 29,596 original artifacts spanning 38 dynasties (2nd c. BCE–19th c. CE), covering 100+ vessel shapes, 200+ glaze colors, and 200+ decorative motifs. To support multimodal model training, we further expand it to over 500K high-quality visual question answering (VQA) pairs via a hybrid pipeline combining expert annotation and LLM-assisted cleaning.
- We establish **CiQi-Bench**, an expert-aligned benchmark for Chinese porcelain connoisseurship, comprising 775 high-quality specimens and two complementary evaluation protocols: (1) a fine-grained multiple-choice setting covering six attributes and standardized naming, and (2) a free-form generation setting assessed via LLM-based attribute-wise similarity scoring.
- We propose a two-phase iterative training framework for CiQi-Agent: Phase I uses GRPO-based RL with a large tool-calling reward to rapidly bootstrap tool-calling skills and generate synthetic trajectories; Phase II integrates these trajectories back into SFT, followed by RL with a reweighted, accuracy-conditioned reward that jointly optimizes tool-calling proficiency and connoisseurship accuracy, thereby aligning CiQi-Agent's tool-calling capability with its domain expertise.
- CiQi-Agent incorporates both visual and retrieval-based tools, including an image zoom-in tool and image/text retrieval tools to enable multimodal RAG. The retrieval database is primarily built from the CiQi-VQA dataset,

consisting of 8,161 curated porcelain pieces with 16,830 images, supplemented by 49,606 cleaned plain-text entries from professional articles.
– The proposed CiQi-Agent, built upon the Qwen2.5-VL-7B-Instruct, achieves superior performance on our benchmark, consistently outperforming all mainstream open- and closed-source multimodal models across all evaluation attributes.

2 Related Work

AI for Porcelain Connoisseurship Early attempts at intelligent porcelain analysis mainly formulate connoisseurship as a fine-grained visual classification problem. Traditional computer vision methods rely on handcrafted texture and color descriptors combined with shallow classifiers for ceramic classification and authentication [25, 32]. With the success of deep learning, convolutional neural networks have significantly improved recognition performance by automatically learning discriminative visual representations for glaze color, vessel shape, and decorative patterns [10, 18]. More recently, multi-task learning has been introduced to jointly recognize multiple porcelain attributes such as dynasty, glaze, ware, and vessel type [15]. Nevertheless, these methods remain recognition-oriented, where the objective is limited to predicting predefined labels from images. They generally lack the capability to perform evidence-grounded reasoning, historical interpretation, or explainable analysis, all of which are fundamental requirements in real-world porcelain connoisseurship.

Tool-Augmented Multimodal Reasoning Recent multimodal large language models (MLLMs), such as BLIP-2 [14] and LLaVA [16], have substantially advanced joint vision-language understanding through instruction tuning and large-scale pretraining. Building upon these models, recent visual agents further enhance reasoning by incorporating external tools, iterative perception, and structured planning. Representative systems employ visual chain-of-thought reasoning [5, 19], program execution [28], image generation [6], or image editing and visual foundation models [33, 35]. Despite their impressive general reasoning capability, these systems are designed for open-domain visual tasks and do not explicitly model domain expertise. Consequently, they are often unable to interpret subtle visual characteristics, retrieve specialized knowledge, or produce historically grounded explanations required for professional artifact connoisseurship.

MLLMs for Cultural Heritage Understanding MLLMs have shown encouraging progress in cultural heritage understanding, including museum exhibit interpretation [2], multilingual artwork understanding [20], cross-temporal cultural reasoning [36], and ancient pottery analysis [7]. These studies demonstrate the potential of vision-language models for assisting cultural heritage analysis by combining visual perception with semantic reasoning. However, existing systems primarily focus on semantic understanding or visual question answering, while expert-level porcelain connoisseurship requires substantially finer-grained visual discrimination, historical attribution, and iterative evidence verification.

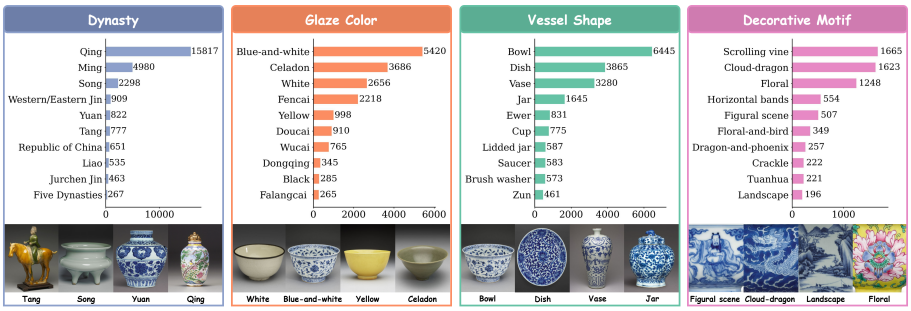

Fig. 2: Visualization of the four key attributes in porcelain connoisseurship. Shown are the distributions of dynasty, glaze color, vessel shape, and decorative motif in the raw porcelain dataset. For each attribute, the top 10 most frequent categories are presented.

3 Dataset and Benchmark

3.1 CiQi-VQA dataset

Raw Data Collection. We first collected raw antique Chinese porcelains from multiple publicly accessible sources, including web-based searches, open-access digital museum collections, and digitized scholarly books. From these sources, we curated a dataset of 29,596 unique specimens spanning 38 dynasties, 42 reign periods, 246 glaze colors, 248 decorative motif categories, and 158 vessel shapes. To the best of our knowledge, this is the most comprehensive dataset for porcelain appreciation currently available.[4] Each porcelain specimen is associated with at least one high-quality image and a standardized name that explicitly encodes **four key attributes**: *dynasty*, *glaze color*, *vessel shape*, and *decorative motif*. In addition, a portion of the specimens further specifies **two additional attributes**—*reign period* and *kiln origin*. A complete example with all six attributes would be: "Qing Dynasty, Kangxi period (1662–1722 CE), Jingdezhen kiln, Blue-and-white, Cloud-and-dragon motif, Bowl (清康熙 景德镇青花云龙纹碗)." Fig. 2 visualizes the distribution across the four key attribute dimensions, showing the top-10 most frequent categories in each.

From this collection, we later selected 775 pieces to construct CiQi-Bench, while the remaining 28,821 pieces were used to build the CiQi-VQA training set.

Metadata Enrichment. We further extracted and cleaned descriptive texts related to porcelain connoisseurship from the raw sources and aligned them with each specimen. However, a substantial portion of the collected porcelain specimens only provided standardized names without detailed narrative descriptions. To address this issue, we invited domain experts to participate in the metadata

[4] A comparison with existing porcelain-related datasets is provided in the supplementary materials.

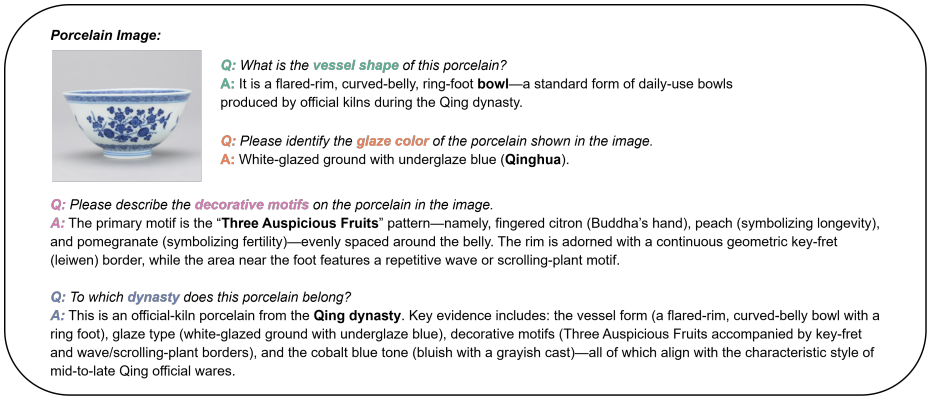

Fig. 3: Examples from CiQi-VQA dataset. The figure illustrates annotations across four key attributes: dynasty, glaze color, decorative motif and vessel shape.

enrichment process. Specifically, for the 61.18% of specimens that lacked detailed descriptions, the expert team composed complementary connoisseurship descriptions based on the specimen images and the available source metadata. In addition, to improve the reliability of the dataset, the experts reviewed and corrected the standardized names of all collected specimens, ensuring that the encoded attribute information was accurate, consistent, and properly formatted. The expert annotation team was led by Honglin Ma from the Shaanxi Academy of Cultural Relics Conservation, a senior researcher with more than 20 years of experience in porcelain identification and connoisseurship research. Under the supervision of Honglin Ma, four graduate students from related disciplines participated in the description completion and standardized-name verification process. Finally, we fed the standardized names, enriched descriptive texts, and raw images into MLLMs to produce a polished connoisseurship description for each specimen, which was structured in six paragraphs corresponding to six connoisseurship attributes.

VQA Data Generation. We leveraged MLLMs to generate specialized VQA pairs targeting the four key attributes—*dynasty*, *glaze color*, *decorative motif*, and *vessel shape*—as illustrated in Fig. 3. We focused on these four rather than the full six benchmark attributes because they are most central to connoisseurship; the remaining two were treated as more advanced refinements and emphasized at the RL stage. For each porcelain specimen, the holistic description was also converted into a VQA format, giving five VQA training samples per item.

To further diversify linguistic expression, we adopted a lightweight augmentation strategy instead of multi-epoch training on identical samples. For each of the five VQA samples, we prompted the LLM to generate four additional variants that preserve the semantics but differ in phrasing and style.

Table 1: Overall statistics of CiQi-VQA dataset and CiQi-Bench.

	Porcelains	Images	Questions		Attributes
			VQA	Multiple-choice	
SFT	28,821	50,675	557,165	—	
RL*	10,275	10,275	10,275	—	dynasty, reign, kiln, color, motif, shape
Evaluation	775	878	775	5,425	
Total	29,596	51,553	557,940	5,425	dynasty, reign, kiln, color, motif, shape

* The raw porcelain data used for RL is a subset of that used for SFT.

The final CiQi-VQA training set thus contains 20 stylistically diverse yet semantically consistent VQA samples per specimen[5], and we performed SFT for a single epoch on this augmented dataset.

3.2 CiQi-Bench

For the benchmark, we curated a set of 775 porcelain specimens and designed two evaluation protocols.

Multiple-Choice Questions. The first protocol adopts a *multiple-choice* format. For each specimen, we constructed seven questions covering the four key attributes, two additional attributes (reign period and kiln origin), and the full standardized name. We used GPT-5 to automatically generate the multiple-choice questions: given the image and the ground-truth annotation as input, the model was instructed to produce plausible yet challenging distractor options. Model performance is then quantified by the answer accuracy over all questions.

Free-Form Questions. The second protocol focuses on *free-form generation*. In this setting, the model is prompted to produce a holistic textual description of each porcelain specimen. An LLM-based evaluator is subsequently employed to compare the generated description with the ground-truth text and assign six separate similarity scores—one for each of the four key attributes and the two additional attributes.[6]

Tab. 1 summarizes the overall statistics of the porcelain dataset and benchmark. The raw porcelain data used for RL is a subset of that used for SFT.

4 Training Framework of CiQi-Agent

In this section, we present the overall architecture of our proposed CiQi-Agent, which emulates the reasoning process of human experts by combining visual perception, retrieval-augmented knowledge access with reinforcement-driven tool-calling. As illustrated in Fig. 4, the agent autonomously analyzes visual details,

[5] For a small subset of specimens (e.g., monochrome wares), only 15 questions are available because these objects inherently lack decorative motifs.

[6] The specific MLLMs and prompt templates used during the construction of our dataset and benchmark are provided in the supplementary materials.

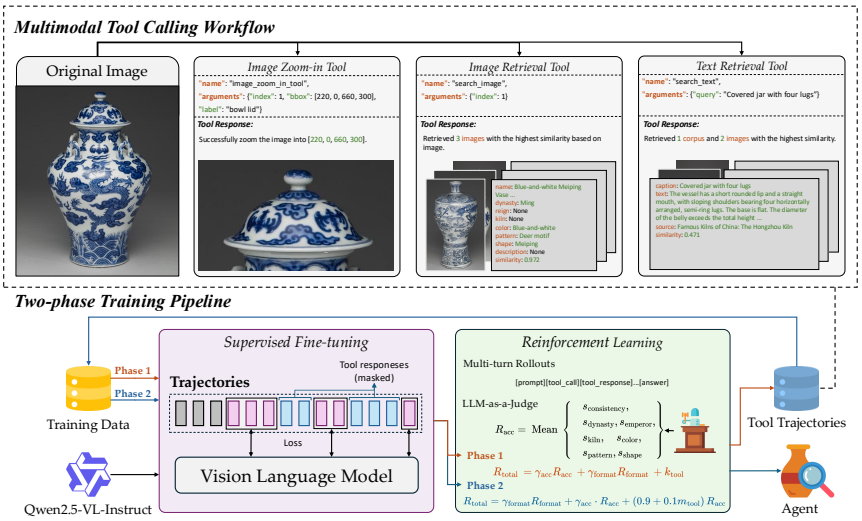

Fig. 4: Framework of the CiQi-Agent. The agent integrates visual zoom-in, image/text retrieval tools within a two-phase training pipeline. Supervised fine-tuning establishes tool-calling skills and porcelain connoisseurship knowledge, while reinforcement learning with an LLM-as-a-Judge refines accuracy and strategic tool-calling.

retrieves multimodal evidence, and generates context-aware judgments through the coordinated use of specialized tools.

4.1 Tool Design

To enable more flexible and interpretable reasoning, a suite of external tools is provided, which can be autonomously selected and executed by CiQi-Agent during inference. Each tool invocation is encapsulated within a standardized `<tool_call></tool_call>` tag pair, allowing the agent to issue structured commands and integrate tool outputs into its reasoning context in a consistent format. These tools are organized into two categories, *vision tool* and *retrieval tool*, which respectively serve the purposes of perceptual enhancement and knowledge acquisition, as illustrated in Fig. 4.

Vision Tool. The *image zoom-in tool* enables the agent to autonomously focus on visually salient regions that are potentially informative for porcelain connoisseurship. Instead of relying on predefined or user-specified coordinates, the agent first analyzes the global image context to infer which local areas merit closer examination—such as decorative motifs, glaze textures, or inscription details. It then dynamically predicts the corresponding bounding-box parameters and extracts high-resolution visual patches from those regions. The resulting sub-images are subsequently reintegrated into the multimodal reasoning context, where they serve as fine-grained perceptual evidence for the ongoing analysis.

Retrieval Tools. Two retrieval tools are provided: an *image retrieval tool* and a *text retrieval tool*, enabling the agent to access evidence from a multimodal porcelain database autonomously. Both tools operate on a unified RAG framework that integrates high-resolution images and textual knowledge related to porcelain connoisseurship.

For *image retrieval*, the query image is encoded by the CLIP encoder [34], and cosine similarity is computed against all image embeddings in the database. The system returns top-k similar entries and their metadata, which are reinserted into the reasoning context as visual evidence. For text retrieval, the query text is encoded by both a CLIP encoder and a text-embedding model [3], yielding two vectors in distinct semantic spaces. Each is matched with its corresponding index via cosine similarity, and the results are fused to identify the most relevant records by a parameter α, which controls the relative contribution of each model to the final retrieval set. The retrieved metadata, such as descriptions, provenance, or linked images, are then supplied to the agent as external knowledge. This dual-space retrieval mechanism grounds the model's reasoning in complementary visual and textual evidence, enhancing factual reliability and interpretability.

4.2 Two-phase Training Pipeline

Given the base model's initial lack of tool-calling competence and the absence of expert-annotated procedural trajectories, the acquisition of tool-calling skills is relegated to unguided exploration in the RL phase. This deficiency leads to a sample-inefficient process that constrains the attainable performance. Consequently, we employ a *two-phase* training pipeline, with each phase consisting of an SFT stage and a subsequent RL stage.

Phase I: Foundational Competence. The first phase aims to establish general connoisseurship knowledge and tool-calling ability. During the SFT, the model is trained on the CiQi-VQA dataset (Sec. 3.1), augmented with instruction-following data [29] to retain general reasoning capability and with 10,575 everyday porcelain samples to reduce overfitting. The subsequent RL stage employs *Group Relative Policy Optimization (GRPO)* [27], and a strong tool-calling reward encourages the model to quickly learn the mechanics of tool-calling, yielding a model that generates synthetic tool-calling trajectories for *Phase II*.

Phase II: Strategic Competence. The second phase focuses on refining the agent's ability to integrate tool-calling with higher-level reasoning. The synthetic trajectories from *Phase I* are merged back into the SFT corpus, allowing the model to internalize both reasoning and tool-calling patterns. The subsequent RL stage employs a redesigned reward that explicitly links tool-calling rewards to connoisseurship accuracy, encouraging the model to improve connoisseurship performance through more effective and purposeful tool-calling. Through this two-phase curriculum, the agent progressively acquires domain-specific expertise required for porcelain connoisseurship while learning to call tools to enhance perception and reasoning.

4.3 Reward Design

During RL, the overall reward function is a weighted average of *format reward* R_{format}, *accuracy reward* R_{acc}, and *tool-calling reward* R_{tool}:

$$R = \gamma_{\text{format}} \cdot R_{\text{format}} + \gamma_{\text{acc}} \cdot R_{\text{acc}} + R_{\text{tool}},$$

where weight parameters γ_{format} and γ_{acc} control the relative importance of the format and accuracy rewards.

Format Reward. The format reward $R_{\text{format}} \in \{0, -1\}$ enforces compliance with the prescribed output and tool-calling syntax. Specifically, the agent receives $R_{\text{format}} = 0$ if both the response and tool calls strictly follow the required format; otherwise, it incurs a penalty of $R_{\text{format}} = -1$.

Accuracy Reward. The accuracy reward R_{acc} measures how accurately the model names the porcelain. It is composed of six attribute scores: *dynasty*, *reign period*, *kiln origin*, *glaze color*, *decorative motif*, and *vessel shape*, plus a *consistency* score that checks whether the output conforms to the required format. Each score $s_i \in [0, 1]$. If the ground-truth does not contain a certain attribute, that attribute is excluded from the average. Formally, let $\mathcal{M}$ denote the set of attribute indices present in the ground-truth for a given sample; then $R_{\text{acc}} = |\mathcal{M}|^{-1} \sum_{i \in \mathcal{M}} s_i$. The evaluation is performed by the LLM-as-a-Judge described in Sec. 4.4.

Tool-calling Reward. The tool-calling reward R_{tool} is designed to guide the agent's learning of tool-calling behavior and its effective use in improving connoisseurship performance.

- During the *first-phase* RL training, the primary objective is to help the model quickly acquire the ability to call tools appropriately. For each rollout, if the model invokes k_{tool} tools, it receives a corresponding reward $R_{\text{tool}} = k_{\text{tool}}$, proportional to the number of successful tool-calls, encouraging early mastery of tool-calling behavior.
- During the *second-phase* RL training, the reward scheme shifts from quantity to quality. To promote strategic and meaningful tool-calling, the agent is rewarded only when the invocation of tools contributes to higher connoisseurship accuracy. Specifically, if the model invokes m_{tool} *distinct* tools during a rollout, it receives a scaled tool-calling reward $R_{\text{tool}} = (0.9 + 0.1 m_{\text{tool}})\, R_{\text{acc}}$ while rollouts without any tool invocation receive no tool-calling reward.

This design incentivizes the model to explore tool-calling strategies that genuinely enhance identification accuracy rather than indiscriminate calling of tools.

4.4 LLM-as-a-Judge

Although porcelain naming follows a standardized attribute structure, linguistic realizations of each attribute exhibit substantial lexical variation while preserving semantic equivalence. As a result, exact string matching or embedding-based similarity metrics fail to reliably assess correctness, particularly when fine-grained, domain-specific terminology is involved. To address this challenge, we

Table 2: Pearson correlation and MAE between human expert scores and LLM-as-a-Judge scores on CiQi-Bench.

	Dynasty	Reign period	Kiln	Glaze color	Motif	Shape
Pearson r	1.00	1.00	0.98	0.96	0.94	0.86
MAE	0.01	0.00	0.04	0.03	0.07	0.08

instead adopt an *LLM-as-a-Judge* approach: during RL training and evaluation, the agent is required to output a standardized naming enclosed within a predefined `<answer></answer>` tag pair. The content inside these tags, along with the corresponding ground-truth porcelain name, is submitted to an LLM-based evaluator. The evaluator is instructed to rate each attribute according to a structured scoring rubric, yielding individual scores $s_i \in [0, 1]$, where higher values indicate greater accuracy or stylistic consistency. The detailed prompt is provided in Sec. 3.4–Sec. 3.5 of the Supplementary Material.

To validate the reliability of LLM-as-a-Judge, we run the SFT model after first-phase training on CiQi-Bench to generate predictions and collect independent evaluations from domain experts using the same scoring rubric. Tab. 2 reports the Pearson correlation coefficients and mean absolute errors (MAE) between expert scores and LLM-based scores across all six attributes. The consistently high correlations, together with the small MAE values, indicate strong alignment between the LLM-based evaluator and human expert judgment.

5 Experiments

In this section, we conduct extensive experiments to evaluate the performance of our proposed CiQi-Agent. We first conduct experiments on the CiQi-Bench, and then perform ablation studies to analyze the impact of each component on our model's performance.

5.1 Experiment Setup

Baseline Configuration. We compare our method against a comprehensive set of state-of-the-art multimodal models, including both closed- and open-source variants. For closed-source models (GPT-5 [22], GPT-4.1 [23], GPT-4o [21], OpenAI o3 [24], Gemini 2.5 Pro [9], and Claude Opus 4 [1]), we use their official APIs with the default inference settings, with the temperature set to 0.0 to ensure deterministic outputs. For open-source models, we select the largest released models for each model family (Qwen2.5-VL-72B-Instruct [26], GLM-4.5V [8], InternVL3.5-241B-A28B-Flash [12], Kimi-VL-A3B-Instruct [13]). Models are loaded using their official implementations and evaluated with the temperature set to 0.0 to ensure reproducibility.

Our Method Configuration. Our agent is built upon Qwen2.5-VL-7B-Instruct, trained using the two-phase pipeline described in Sec. 4, with Qwen2.5-72B-Instruct [30] served as the LLM-as-a-Judge evaluator. The weight parameters are set as $\gamma_{\text{format}} = 0.2$ and $\gamma_{\text{acc}} = 1.0$, respectively. For retrieval, we construct a database consisting of 16,830 images and 49,606 texts from our CiQi-VQA dataset, which is strictly non-overlapping with both the RL training set and the CiQi-Bench to avoid information leakage and ensure fair assessment. See Sec. 4 of the Supplementary Material for the detailed configurations.

Evaluation Metrics. We report accuracy as the primary metric for multiple-choice questions, computed as the percentage of correctly answered questions across all seven dimensions: overall naming, dynasty, reign period, kiln site, glaze color, decorative motif, and vessel shape. For free-form generation tasks, we employ Qwen2.5-72B-Instruct [30] as the LLM-as-a-Judge evaluator.

5.2 Main Results on CiQi-Bench

Tab. 3 summarizes the results on CiQi-Bench. CiQi-Agent achieves **state-of-the-art performance** across all dimensions, validating the effectiveness of our two-phase training pipeline and tool-augmented reasoning framework.

On multiple-choice tasks, CiQi-Agent attains 85.2% overall naming accuracy and an average of 81.5% across all seven attributes, surpassing GPT-5 (average 75.8%) and Qwen2.5-VL-72B-Instruct (average 69.2%). Notably, it achieves 77.6% on dynasty, 70.3% on reign period, and 81.8% on kiln site—outperforming all baselines by substantial margins. Performance gains are especially pronounced in visually grounded attributes, reaching 91.4% on glaze color, 75.7% on decorative motif, and 88.1% on vessel shape.

For free-form generation, CiQi-Agent further demonstrates strong generalization, achieving 71.3% on dynasty (vs. 42.7% for o3), 69.8% on kiln site (vs. 44.4% for o3), and 85.4% on glaze color (vs. 75.8% for Qwen2.5-VL-72B), with an average of 66.7% across all six attributes (vs. 43.0% for Qwen2.5-VL-72B and 48.0% for GPT-5). Remarkably, despite having only 7B parameters, CiQi-Agent outperforms GPT-5 across all attributes and in both multiple-choice and free-form averages, demonstrating that tool-calling and domain-aligned training can effectively compensate for model scale.

5.3 Ablation Study

Training stages. To understand how each training stage contributes to the final performance, we perform an ablation study based on Qwen2.5-VL-7B-Instruct, as shown in Tab. 4.

We conduct an ablation study with the following three model variants: (1) the base model; (2) the SFT model fine-tuned on the base model, corresponding to the *Phase II* model; (3) the SFT+RL model, i.e., our full **CiQi-Agent**. Since the tool-calling trajectories used for RL training are generated by the *Phase I* model, Model (3) is built on Model (2).

Table 3: Performance on CiQi-Bench (Multiple-choice & Free-form). Bold indicates the best performance, and underline indicates the second-best performance.

Model	Multiple-choice Accuracy (%)								Free-form Accuracy (%)						
	Dynasty	Reign	Kiln	Color	Motif	Shape	Naming	Average	Dynasty	Reign	Kiln	Color	Motif	Shape	Average
GPT-5 [22]	<u>65.7</u>	61.4	<u>79.6</u>	86.5	69.3	83.8	<u>84.3</u>	<u>75.8</u>	39.4	32.8	42.6	74.4	<u>35.3</u>	63.9	48.0
GPT-4.1 [23]	59.3	<u>68.3</u>	71.1	85.0	62.2	81.8	77.9	72.2	36.7	27.2	29.0	67.5	27.6	60.1	41.3
GPT-4o [21]	59.1	60.4	68.6	<u>89.2</u>	70.1	<u>84.2</u>	82.1	73.4	26.9	13.4	15.1	53.9	21.1	47.6	29.7
o3 [24]	57.6	57.4	72.2	82.6	62.4	76.8	76.6	69.4	<u>42.7</u>	<u>36.6</u>	44.4	74.2	33.1	62.1	<u>48.8</u>
Gemini 2.5 Pro [9]	54.4	57.4	68.0	65.2	58.5	64.2	82.5	64.3	<u>48.1</u>	34.9	<u>48.4</u>	50.2	16.1	39.0	39.5
Claude Opus 4 [1]	54.8	40.6	65.3	74.9	59.0	75.2	69.3	62.7	36.8	10.3	22.0	64.1	25.1	59.1	36.2
Qwen2.5-VL-72B-Instruct [26]	57.6	34.7	69.2	86.7	<u>71.7</u>	84.1	80.3	69.2	29.5	31.2	27.7	<u>75.8</u>	31.0	62.6	43.0
GLM-4.5V (106B) [8]	58.3	59.4	75.8	82.3	70.4	81.8	80.6	72.6	31.0	14.3	32.8	65.4	31.1	<u>65.2</u>	39.9
InternVL3.5-241B-A28B-Flash [12]	57.1	38.6	59.5	82.1	64.8	73.9	68.5	63.5	42.4	31.6	36.9	52.6	19.6	41.5	37.4
Kimi-VL-A3B-Instruct (16B) [13]	59.3	22.8	48.8	84.8	59.8	77.9	70.3	60.5	17.3	23.7	16.2	69.5	26.5	61.3	35.7
CiQi-Agent (Ours, 7B)	**77.6**	**70.3**	**81.8**	**91.4**	**75.7**	**88.1**	**85.2**	**81.5**	**71.3**	**49.1**	**69.8**	**85.4**	**49.7**	**75.0**	**66.7**

Table 4: Ablation study on Qwen2.5-VL-7B-Instruct variants. Arrows indicate improvement (↑) or decline (↓) relative to the previous variant, and bold indicates the best performance.

Model	Multiple-choice Accuracy (%)								Free-form Accuracy (%)						
	Dynasty	Reign	Kiln	Color	Motif	Shape	Naming	Average	Dynasty	Reign	Kiln	Color	Motif	Shape	Average
Qwen2.5-VL-7B-Instruct [26]	54.0	60.4	55.1	83.6	69.8	79.6	69.0	67.4	20.3	22.2	15.5	70.6	28.4	61.4	36.4
+ SFT	65.5 ↑	53.5 ↓	77.1 ↑	**92.6** ↑	72.8 ↑	**88.2** ↑	81.9 ↑	75.9 ↑	64.6 ↑	45.2 ↑	56.6 ↑	81.6 ↑	35.8 ↑	70.9 ↑	59.1 ↑
+ SFT + RL	**77.6** ↑	**70.3** ↑	**81.8** ↑	91.4 ↓	**75.7** ↑	88.1 ↓	**85.2** ↑	**81.5** ↑	**71.3** ↑	**49.1** ↑	**69.8** ↑	**85.4** ↑	**49.7** ↑	**75.0** ↑	**66.7** ↑

Effect of SFT. SFT substantially boosts performance: overall naming improves from 69.0% to 81.9%, and the multiple-choice average rises from 67.4% to 75.9%. On free-form evaluation, dynasty accuracy jumps from 20.3% to 64.6%, and the average increases from 36.4% to 59.1%, establishing core connoisseurship knowledge.

Effect of RL. Adding RL aligns tool-calling with knowledge-grounded reasoning: kiln accuracy jumps from 67.2% to 81.8% and dynasty from 51.6% to 77.6%, and all free-form attributes improve (e.g., kiln: 63.6% → 69.8%, dynasty: 66.8% → 71.3%). Correspondingly, the multiple-choice average increases from 69.7% to 81.5%, and the free-form average from 62.3% to 66.7%, yielding our final CiQi-Agent.

Tool Configurations. To understand how each tool contributes to the final performance, we perform an ablation study on GPT-5, Qwen2.5-VL-7B-Instruct, and our CiQi-Agent, as shown in Tab. 5.

Tool on Base Models. Tools do not consistently benefit general-purpose base models. (e.g., GPT-5's average: 75.8% → 76.3%/70.8%/70.8%, Qwen2.5-VL-7B-Instruct's average: 67.4% → 47.1%/48.6%/48.1%). A plausible explanation is that the returned evidence (retrieved text or visual cues) is domain-specific and may require porcelain-appraisal knowledge to interpret and reconcile, which these base models do not reliably exhibit. In contrast, CiQi-Agent shows consistent gains (e.g., average: 69.7%/75.1%/77.3% → 81.5%), suggesting that models adapted with porcelain-domain expertise can more effectively interpret tool-provided evidence and translate it into improved performance.

Table 5: Ablation study on GPT-5, Qwen2.5-VL-7B-Instruct and CiQi-Agent with different tool configurations. Bold indicates the best performance.

Model	Multiple-choice Accuracy (%)							
	Dynasty	Reign	Kiln	Color	Motif	Shape	Naming	Average
GPT-5 [22]	65.7	61.4	79.6	86.5	69.3	83.8	84.3	75.8
+ vision tool	62.4	68.3	80.4	86.9	67.7	83.2	85.0	76.3
+ retrieval tools	54.9	58.4	69.4	85.5	65.6	81.4	80.2	70.8
+ all tools	55.7	58.4	68.3	85.8	66.7	82.3	78.6	70.8
Qwen2.5-VL-7B-Instruct [26]	54.0	60.4	55.1	83.6	69.8	79.6	69.0	67.4
+ vision tool	32.7	51.5	31.4	55.0	44.4	54.7	60.1	47.1
+ retrieval tools	33.6	54.5	35.5	55.4	45.2	55.6	60.6	48.6
+ all tools	32.3	53.5	36.9	54.1	45.2	55.4	59.3	48.1
CiQi-Agent (Ours, 7B) (w/o tools)	51.6	32.7	67.2	**92.7**	73.3	86.2	84.3	69.7
+ vision tool	69.2	66.3	59.8	91.0	**78.6**	76.7	84.2	75.1
+ retrieval tools	68.7	59.4	81.0	88.9	75.4	83.3	84.3	77.3
+ all tools	**77.6**	**70.3**	**81.8**	91.4	75.7	**88.1**	**85.2**	**81.5**

Tool Combination on CiQi-Agent. Combining tools is more effective than using a single tool for CiQi-Agent. The best average performance is achieved when vision and retrieval tools are enabled together (average: 81.5%), which supports the view that the two tools provide complementary information that is most useful when integrated.

Overall, the ablation results reveal a clear division of labor among training stages and tools. SFT provides a comprehensive uplift of the model's domain knowledge and yields consistent improvements across all evaluations. Building on this foundation, RL with the vision tool strengthens the model's ability to capture fine-grained visual cues, which is particularly beneficial for visually grounded attributes such as motif. When multimodal retrieval tools are further incorporated, the agent gains the ability to compare input images against external porcelain exemplars, significantly improving history-based attributions, including dynasty, reign, and shape.

6 Conclusion

In this work, we present CiQi-Agent, a domain-specific multimodal agent for antique Chinese porcelain connoisseurship. We build CiQi-VQA, a large-scale dataset of expert-curated porcelain images and question–answer pairs, and CiQi-Bench, an expert-aligned benchmark that evaluates six connoisseurship attributes. For the training framework, we design a two-phase training pipeline that combines SFT, RL, and tool-augmented reasoning to align tool-calling with domain expertise. CiQi-Agent integrates visual zoom-in and image/text retrieval tools to perform fine-grained analysis with multimodal RAG. Extensive experiments show that CiQi-Agent significantly outperforms mainstream MLLMs across all

attributes, demonstrating the effectiveness of our dataset, benchmark, and training framework for cultural-heritage connoisseurship.

For future work, we plan to move beyond connoisseurship and tackle more challenging tasks of authentication, i.e., distinguishing genuine antique porcelains from later imitations. In addition, CiQi-Agent represents a first step toward using MLLMs for cultural-heritage analysis; the same framework can be extended to build agents for other artifact types (e.g., ancient coins, calligraphy, paintings) or to develop a more general foundation model for cultural-heritage connoisseurship.

Acknowledgments

This work is supported by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant No. 25GA3200200, and the Shanghai Artificial Intelligence Laboratory. We gratefully acknowledge the valuable support from the Shaanxi Academy of Cultural Relics Conservation. In particular, we sincerely thank expert Hongli Ma and his team for their valuable guidance and support in the construction of our dataset.

References

1. Anthropic: Claude 4 system card. Tech. rep., Anthropic (2025), https://www.anthropic.com/claude-4-system-card
2. Balauca, A.A., Garai, S., Balauca, S., Shetty, R.U., Agrawal, N., Shah, D.S., Fu, Y., Wang, X., Toutanova, K., Paudel, D.P., Van Gool, L.: Understanding museum exhibits using vision-language reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2227–2238 (October 2025)
3. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216 (2024)
4. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24185–24198 (2024). https://doi.org/10.1109/CVPR52733.2024.02283
5. Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Sun, Z., Gutfreund, D., Gan, C.: Visual chain-of-thought prompting for knowledge-based visual reasoning. In: AAAI. vol. 38, pp. 1254–1262 (2024)
6. Chern, E., Hu, Z., Chern, S., Kou, S., Su, J., Ma, Y., Deng, Z., Liu, P.: Thinking with generated images. arXiv preprint arXiv:2505.22525 (2025)
7. Ge, J., Cheng, T., Wu, B., Zhang, Z., Huang, S., Bishop, J., Shepherd, G., Fang, M., Chen, L., Zhao, Y.: Vasevqa: Multimodal agent and benchmark for ancient greek pottery. arXiv preprint arXiv:2509.17191 (2025), concurrent work
8. GLM-V Team: Glm-4.1v-thinking and glm-4.5v: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
9. Google DeepMind: Gemini 2.5 pro (2025)

10. Hu, Y., Wu, S., Ma, Z., Cheng, S.: Integrating deep learning and machine learning for ceramic artifact classification and market value prediction. npj Heritage Science **13**(1), 1–17 (2025)
11. Huang, Y., Sheng, X., Yang, Z., Yuan, Q., Duan, Z., Chen, P., Li, L., Lin, W., Shi, G.: Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 5911–5920. MM '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664647.3680649
12. InternVL Team: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
13. Kimi Team: Kimi-VL technical report. arXiv preprint arXiv:2504.07491 (2025)
14. Li, J., Li, D., Savarese, S., Hoi, S.C.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning (ICML) (2023)
15. Ling, Z., Delnevo, G., Salomoni, P., Mirri, S.: Multi-task learning for identification of porcelain in song and yuan dynasties. arXiv preprint arXiv:2503.14231 (2025)
16. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning for large language and vision assistant. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
17. Liu, Y., Zhou, L., Zhang, P., Bai, X., Gu, L., Yu, X., Zhou, J., Hancock, E.R.: Where to focus: Investigating hierarchical attention relationship for fine-grained visual classification. In: ECCV. pp. 57–73 (2022)
18. Liu, Y., Liu, B., Yu, J., Xia, J., Luo, C.: Automatic classification of blue and white porcelain sherds based on data augmentation and feature fusion. Applied Artificial Intelligence **36**(1), 1994232 (2022)
19. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain-of-thought prompting for large multimodal models. In: CVPR. pp. 14420–14431 (June 2024)
20. Mohamed, Y., Li, R., Ahmad, I.S., Haydarov, K., Torr, P., Church, K., Elhoseiny, M.: No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 20939–20962. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.1165
21. OpenAI: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
22. OpenAI: GPT-5 system card. Tech. rep., OpenAI (2025)
23. OpenAI: Introducing GPT-4.1 in the api (2025)
24. OpenAI: Introducing openai o3 and o4-mini (Apr 2025)
25. Pavithra, L.K., Sree Sharmila, T., Subbulakshmi, P.: Texture image classification and retrieval using multi-resolution radial gradient binary pattern. Applied Artificial Intelligence **35**(15), 2298–2326 (2021)
26. Qwen Team: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
27. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
28. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. In: ICCV. pp. 11888–11898 (October 2023)
29. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model (2023)
30. Team, Q.: Qwen2.5: A party of foundation models (September 2024)

31. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution (2024)
32. Weng, Z., Guan, Y., Luo, H.: Machine vision based classification and identification for non-destructive authentication of ancient ceramic. Journal of the Chinese Ceramic Society **45**(12), 1833–1842 (2017)
33. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
34. Yang, A., Pan, J., Lin, J., Men, R., Zhang, Y., Zhou, J., Zhou, C.: Chinese clip: Contrastive vision-language pretraining in chinese (2023)
35. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
36. Zhou, L., Yu, L., Xie, D., Cheng, S., Li, W., Li, H.: Hanfu-bench: A multimodal benchmark on cross-temporal cultural understanding and transcreation. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 24627–24649. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.1251