A Structured Framework for Evaluating and Enhancing Interpretive Capabilities of Multimodal LLMs in Culturally Situated Tasks

Anonymous ACL submission

Abstract

This study aims to test and evaluate the capabilities and characteristics of current mainstream Visual Language Models (VLMs) in generating critiques for traditional Chinese painting. To achieve this, we first developed a quantitative framework for Chinese painting critique. This framework was constructed by extracting multi-dimensional evaluative features—including evaluative stance, core focal points, and argumentative quality—from human expert critiques using a zero-shot classification model. Based on these features, several representative critic personas were defined and quantified. This framework was then employed to evaluate selected VLMs (e.g., Gemini 2.5 Pro). The experimental design involved persona-guided prompting to assess the VLM's ability to generate critiques from diverse perspectives. Our findings reveal the current performance levels, strengths, and areas for improvement of VLMs in the domain of art critique, offering insights into their potential and limitations in complex semantic understanding and content generation tasks. The code used for our experiments can be publicly accessed at: https://github.com/anon_user/anon_repo[1].

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance on general NLP benchmarks, yet their applicability in culturally embedded, humanistic domains remains limited. In high-context interpretive tasks such as art criticism, clinical narrative analysis, or historical commentary, model performance depends not only on linguistic fluency or factual accuracy, but also on deeper forms of cognitive alignment—epistemic sensitivity, rhetorical coherence, and cultural adaptability.

A representative and particularly demanding testbed for such capabilities is *Chinese art commentary*. This genre, especially when analyzing works like traditional landscape or court paintings, involves symbolic interpretation, aesthetic judgment, and deeply situated cultural discourse. Existing multimodal LLMs are rarely evaluated in this space. Standard benchmarks such as MME (Fu et al., 2024) and MMBench (Liu et al., 2024) focus on object recognition or task-oriented vision-language reasoning, while frameworks like ArtGPT (Chen et al., 2024) emphasize captioning and factual grounding. These methods largely overlook interpretive nuance and disciplinary diversity.

Meanwhile, humanistic commentary often exhibits non-linear logic, specialized lexicons, and varied stylistic conventions, particularly in Chinese art contexts where rhetorical strategies such as *yijing* (意境, artistic conception) or *qiyun shengdong* (气韵生动, spiritual resonance) are essential but difficult to quantify. Without appropriate grounding, LLMs risk producing synthetic outputs that mimic surface patterns but fail to demonstrate epistemic alignment (Guo et al., 2023; Mishra et al., 2024). This growing mismatch calls for new paradigms in evaluation and adaptation.

To address these challenges, we introduce VULCA—the *Vision-Understanding and Language-based Cultural Adaptability Framework*. VULCA is a structured evaluation and enhancement framework designed to assess how well MLLMs align with domain-specific interpretive practices in culturally situated tasks. Our work centers on Chinese art commentary, but the methodology generalizes to other multimodal and epistemically rich domains such as religion, medicine, or history. VULCA combines three core components: (1) a multi-dimensional human expert benchmark (MHEB) constructed from 163

[1]Repository will be linked upon paper acceptance.

art commentaries annotated across five cultural capability dimensions; (2) a persona-guided recontextualization mechanism using eight interpretive personas and a domain-specific knowledge base; and (3) a joint evaluation pipeline integrating vector-space semantic alignment with rubric-based capability scoring. Commentaries are generated from annotated traditional Chinese paintings, and their alignment with expert patterns is evaluated with and without interventions. As a result, we produce five contributions: (i) the definition of VULCA, a new structured framework for assessing and enhancing MLLMs in culturally grounded, multimodal reasoning tasks; (ii) we construct D1, a high-quality human benchmark of Chinese art commentary annotated across five capability dimensions; (iii) we develop and evaluate persona-guided recontextualization interventions using eight expert personas and a domain-specific knowledge base; (iv) we demonstrate over 20% improvement in symbolic reasoning and over 30% improvement in argumentative coherence on Gemini 2.5 Pro using our proposed method; and (v) we establish the generalizability of our evaluation methodology to other epistemically rich domains such as religion, history, and education.

Together, our work highlights the need for new evaluation paradigms that go beyond benchmark metrics and toward measuring how well LLMs can adapt to the interpretive demands of real-world, interdisciplinary contexts.

2 Related Work

Missing Evaluation Dimensions for Cultural Reasoning. Existing benchmarks for large or multimodal language models, such as (Fu et al., 2024; Li et al., 2023), emphasize factual accuracy or instruction following, seldom addressing symbolic interpretation or epistemic alignment. ArtGPT (Chen et al., 2024), for instance, evaluates stylistic generation but lacks formal metrics for interpretive depth. While prior work explores aesthetic reasoning (Qi, 2024; Wang, 2024), these studies rarely offer structured, multi-capability evaluation. Our work introduces cultural adaptability, operationalized through a multi-dimensional human expert benchmark (MHEB) with capability rubrics, enabling quantitative comparison in high-context domains like Chinese art.

Limitations of Persona Conditioning Without Grounding. Persona use in LLM evaluation shows promise for style control (Wang et al., 2023a, 2024), yet most methods lack structured knowledge grounding, especially in epistemically rich domains. Our method combines persona simulation with curated domain-specific knowledge to guide generation towards symbolic reasoning and cultural interpretation, not just stylistic alignment, offering a controlled intervention mechanism.

Gap in Multimodal Input–Interpretation Evaluation. Current multimodal frameworks like MMBench or LLaVA-Bench (Zhang et al., 2023) primarily focus on classification, question answering, or instruction following, rarely requiring grounded interpretation. Our pipeline links annotated symbolic elements with structured prompts for art commentary, evaluating MLLM outputs for semantic alignment with MHEB using vector-space and rubric-based metrics, addressing a gap in assessing image-conditioned cultural reasoning.

Lack of Comparative Cultural Interventions Across Models. Surveys (Mishra et al., 2024; Guo et al., 2023) discuss LLM limitations in nuanced discourse, but few studies compare model responsiveness to structured cultural interventions. Our empirical evaluation shows persona and knowledge base intervention improves symbolic reasoning and argumentative coherence by over 20–30%, highlighting epistemic alignment's role beyond fluency. This cross-model, capability-specific analysis distinguishes our work.

3 Methodology

This research aims to comprehensively evaluate Visual Language Models (VLMs) capabilities in generating critiques for traditional Chinese painting, assessing their understanding of image content, commentary quality, and adaptability to guided perspectives. The workflow involves: Framework Construction, developing a quantitative analytical framework from human expert commentaries, including defining evaluative dimensions and critic personas; VLM Evaluation Experiment Design, creating structured protocols for VLM critique generation

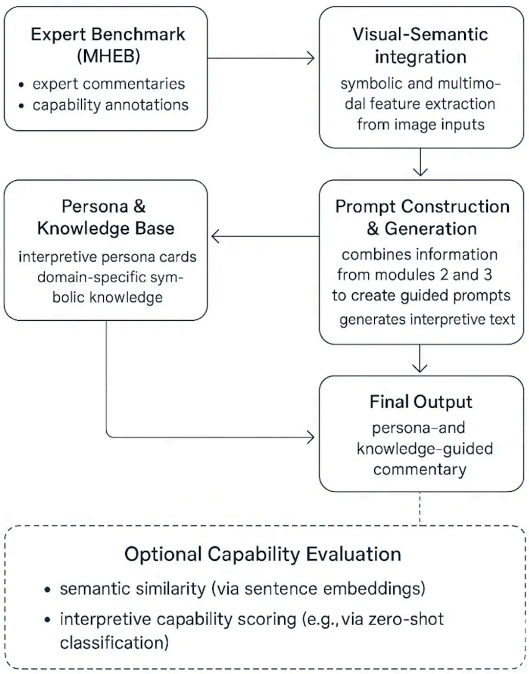

Figure 1: Overview of the VULCA framework, illustrating its components and their interactions for structured evaluation and intervention in art criticism.

under conditions like persona-based and baseline prompting; and Experimentation and Result Analysis, implementing experiments, collecting VLM critiques, and analyzing them with the developed framework to assess capabilities and intervention impacts. Figure 1 provides an overview of this framework and its components.

A cornerstone is the quantitative framework benchmark for VLM critiques, built upon human expert commentaries on Chinese art. To ensure objective, reproducible, and fine-grained evaluation, an automated capability assessment framework was developed. This involves feature extraction, multi-dimensional capability scoring, profile assignment, and visualization, using a zero-shot classification model for fine-grained evaluative labels. The scoring covers painting element recognition, Chinese painting understanding, and language usage, each with a dedicated rubric . This structured, rule-based approach enhances objectivity and facilitates large-scale benchmarking (Jiang and Chen, 2025; Hayashi et al., 2024).

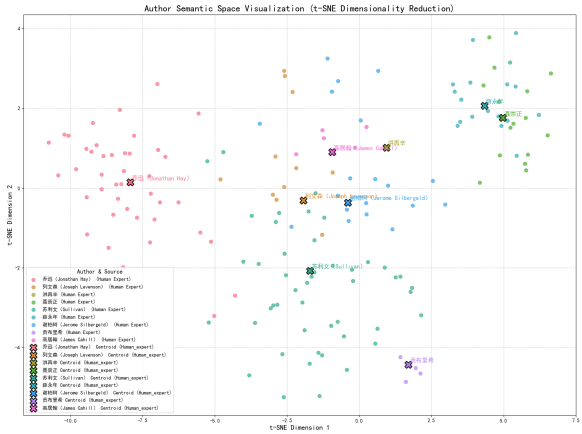

Figure 2: T-SNE visual representation of human expert art commentaries.

3.1 Feature Engineering from Human Expert Critiques

Framework foundation relies on human expert commentaries, significantly from Giuseppe Castiglione's (Lang Shining) "Twelve Months" (十二月令图) series—Qing imperial court paintings fusing Chinese and Western traditions. To enhance model training and evaluation, a sliding window cropping strategy (640×640 pixel sub-images) was applied to these high-resolution images, augmenting data diversity and granularity for improved VLM detail recognition and evaluation accuracy, a common practice in computer vision (e.g., (Lin et al., 2014; Krishna et al., 2017)).

We employed a zero-shot classification model to systematically extracted evaluative characteristics. This model, proficient in multilingual text classification without task-specific training, objectively identified and scored texts against predefined labels across three dimensions: *Evaluative Stance* (e.g., "Historical Research"), *Core Focal Points* (e.g., "Use of Color"), and *Argumentative Quality* (e.g., "Profound Insight"). This process created a structured, multi-dimensional feature representation for each expert commentary. Appendix D.5 lists these labels. Figure 2 visualizes the MHEB semantic distribution from these features.

The zero-shot classification model serves as an analytical tool for deconstructing expert texts and building our evaluation framework, distinct from the VLMs (e.g., Gemini 2.5 Pro, Qwen-VL) evaluated later.

3.2 Evaluation Dimensions and Label System

To systematically analyze commentary content, we define a structured annotation scheme based on three major dimensions: **Stance**, **Feature Focus**, and **Commentary Quality**. These dimensions were derived from domain-specific literature and refined through pilot studies with expert annotations.

Stance characterizes the rhetorical or evaluative position taken by the commentator (e.g., historical interpretation, praise, or critique). Feature Focus identifies the specific visual or contextual aspects discussed in the commentary (e.g., line quality, symbolism, spatial composition). Commentary Quality captures the analytical depth and logical structure of the commentary, ranging from clear, well-argued insights to superficial or biased remarks.

Each dimension comprises a set of fine-grained subcategories with bilingual English–Chinese mappings. Full definitions and label lists are provided in Appendix D.5.

To illustrate the system, we briefly explain one representative label from each dimension:

Stance –Aesthetic Appreciation (美学鉴赏型): Commentary focuses on the beauty and expressive power of the painting, often using evocative or poetic language to highlight visual elegance or emotional resonance. **Feature Focus –Brushwork Technique (笔法技巧):** The analysis emphasizes the artist's brushstroke styles, control, or variation, such as fine lines, dry brush texture, or fluid ink application. **Commentary Quality –Profound Insight (见解深刻独到):** The argument demonstrates deep understanding, originality, and relevance, going beyond surface observations to offer meaningful interpretations.

These representative sub-dimensions help bridge formal annotation with art historical reasoning. The full taxonomy serves as the foundation for profile construction, persona classification, and performance evaluation.

3.3 Construction and Definition of Critic Personas

To capture holistic critique style and depth beyond granular features, we constructed "critic personas" representing archetypal critical perspectives. Their development was data-driven, analyzing features from human expert commentaries, complemented by art history domain expertise. Five core personas were defined: Comprehensive Analyst (博学通论型), Historically Focused Critic (历史考据型), Technique & Style Focused Critic (技艺风格型), Theory & Comparison Focused Critic (理论比较型), and General Descriptive Profile (泛化描述型).

Each persona is quantitatively defined by rules and thresholds based on zero-shot classification feature scores . This rule-based matching objectively assigns commentaries (human or VLM) to personas. Persona definition and matching rely on explicit features and rule-based logic, not primarily direct semantic embedding of raw text. Dimensionality reduction (t-SNE/UMAP) visualizes commentary and persona distribution in the feature space, not for initial persona vector generation.

3.4 Value and Application of the Framework

The resulting quantitative framework, which integrates fine-grained feature analysis with the abstracted critic personas, offers a multi-dimensional, quantifiable, and empirically grounded benchmark. Rooted in the discernible characteristics of human expert critiques, this framework provides a structured and robust foundation for the subsequent systematic evaluation and comparative analysis of Chinese painting commentaries generated by Visual Language Models.

3.5 Experimental Design for VLM Evaluation

This quantitative framework guided experiments evaluating selected VLMs (e.g., Gemini 2.5 Pro, Qwen-VL). The core task required VLMs to generate commentary on provided traditional Chinese painting images. Experiments typically involved structured, multi-round interactions for each VLM per image, including persona-based and baseline Q&A rounds (Zhou et al., 2024; Wang et al., 2023b).

Inputs were multifaceted: high-definition "Monthly Images" (sometimes segmented); predefined "Persona Cards" (Wang et al., 2023b) guiding analysis—Mama Zola (佐拉妈妈), Professor Elena Petrova (埃琳娜·佩特洛娃教授), Okakura Kakuzō (冈仓天心), Brother Thomas (托马斯修士), John Ruskin (约翰·罗斯金), Su Shi (苏轼), Guo Xi (郭熙), Dr. Aris Thorne (阿里斯·索恩博士); standardized prompt templates

(Zhou et al., 2024); and an optional JSON knowledge base (Zhang et al., 2024b; Bin et al., 2024). Persona guidance aimed to assess VLM capability to simulate diverse perspectives and analytical styles (Zhang et al., 2024a).

All VLM-generated texts were recorded and systematically organized. These outputs were then analyzed using the quantitative framework (Section 3.2), applying zero-shot classification to extract feature scores and matching critiques against predefined ”Critic Personas” to assess alignment, especially under specific persona guidance.

Key VLM evaluation dimensions include: *Painting Element Recognition* (5-point scale); *Chinese Painting Understanding* (7-point scale); and *Chinese Language Usage* (5-point scale). Prompt design, particularly for structured commentary, targeted these dimensions .

3.6 Vector Space Representation and Visualization

To compare human and VLM critiques, we converted feature scores (evaluative stance, focal points, argumentative quality) from both into numerical vectors. These vectors were projected into a 2D space using t-SNE for visualisation (Reimers and Gurevych, 2019), enabling assessment of semantic similarity and distributional differences. Figure 2 illustrates such a comparative visualisation, showing the semantic distribution of human expert commentaries versus baseline MLLM-generated commentaries, highlighting their initial semantic gap.

These visualizations help analyze how MLLM outputs align with human expert benchmarks, identify specific MLLM strengths/weaknesses, and assess persona/knowledge interventions' impact on aligning MLLM critiques with desired expert profiles.

3.7 Automated Workflow

This research implemented a modular, automated experimental pipeline for profile scoring, dimensionality reduction, and dataset preparation for visualizations.

Experimental benchmarking involved MLLM commentary generation using a curated artwork dataset with varied prompts (baseline, persona-specific, knowledge-enhanced). MLLM outputs were logged, versioned, and organized by model, persona, and prompt. Subsequent automated analysis involved feature extraction, persona scoring, and comparative metrics generation. This systematic approach facilitated large-scale, reproducible evaluation of MLLM performance in Chinese art critique.

3.8 Multi-Model Comparative Evaluation

To comprehensively assess the capabilities of state-of-the-art large language and vision-language models, we conducted a systematic comparative evaluation across four representative models: Google Gemini 2.5 Pro, Meta Llama-3.1-8B-Instruct, Meta Llama-4-Scout-17B-16E-Instruct, and Qwen-2.5-VL-7B. All models were evaluated using the same experimental protocol, dataset splits, and evaluation metrics to ensure fair and reproducible comparison.

All models were accessed via their official APIs or open-source checkpoints, with inference settings kept consistent. For multimodal tasks, only models supporting both text and image inputs were included in the corresponding benchmarks. The evaluation covers a range of tasks, including argumentative quality, core focal points, stance analysis, and semantic space visualization, as detailed in Section D.

3.9 Quantitative Modeling and Formalisms

This section details the key mathematical formulations used in our analytical framework, covering semantic representation, comparative metrics, and the profile matching algorithm.

Semantic Embedding. Conceptually:

$$\mathbf{v}_d = \text{SentenceTransformer}(\text{document}_d) \quad (1)$$

Where ($\mathbf{v}_d \in \mathbb{R}^N$) (e.g., ($N = 1024$) for BAAI/bge-large-zh-v1.5).

Average Quality Score for Radar Chart ($\bar{q}_{j,G}$). For a quality dimension j and a group of documents G (e.g., Human Experts, MLLM Baseline):

$$\bar{q}_{j,G} = \frac{1}{|N_G|} \sum_{d \in N_G} s_{j,d} \quad (2)$$

Where $s_{j,d}$ is the score of document d on quality dimension j, and $|N_G|$ is the number of documents in group G.

Centroid Calculation in Dimensionality Reduced Space ($\mathbf{c}_p$). For a profile/condition p, its centroid in a 2D space (e.g., t-SNE):

$$\begin{aligned}\mathbf{c}_p &= (\bar{x}_p, \bar{y}_p) \\ &= \left(\frac{1}{|D_p|} \sum_{d \in D_p} x_d, \frac{1}{|D_p|} \sum_{d \in D_p} y_d \right)\end{aligned} \tag{3}$$

Where (x_d, y_d) are the 2D coordinates of document d belonging to profile/condition p, and $|D_p|$ is the number of documents in profile/condition p.

Cohen's d (Effect Size). To measure the standardized difference between two group means $(\bar{X}_1, \bar{X}_2)$:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \tag{4}$$

Where s_p is the pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{5}$$

And here n_1, n_2 are the sample sizes of group 1 and group 2, while s_1^2, s_2^2 are the variances of group 1 and group 2.

Stance Contribution Formula (S_C). We compute the stance contribution S_C using the following conditions:

$$S_C = \begin{cases} \frac{s_{\text{actual}} - s_{\text{min\_rule}}}{s_{\text{max\_rule}} - s_{\text{min\_rule}}}, & \text{if } L_{\text{actual}} = L_{\text{rule}}, \\ & s_{\text{actual}} \geq s_{\text{min\_rule}}, \\ & s_{\text{max\_rule}} \neq s_{\text{min\_rule}} \\ 1, & \text{if } L_{\text{actual}} = L_{\text{rule}}, \\ & s_{\text{actual}} \geq s_{\text{min\_rule}}, \\ & s_{\text{max\_rule}} = s_{\text{min\_rule}} \\ 0, & \text{otherwise} \end{cases}$$

Where S_C is the stance contribution score, L_{actual} is the actual stance label of the text, L_{rule} is the required stance label in the profile rule, s_{actual} is the actual stance score, and $s_{\text{min\_rule}}$, $s_{\text{max\_rule}}$ represent the required range.

4 Results

We present results from semantic alignment, capability profiling, and the effects of persona-guided interventions on MLLMs. All evaluations are made with respect to the multi-dimensional human expert benchmark (MHEB), using both vector-space analysis and rubric-based scoring.

4.1 Semantic Divergence from Expert Commentary

Baseline MLLM outputs exhibit significant divergence from human expert commentaries. As shown in Figure 3 (left), expert texts cluster tightly in semantic space, while MLLM outputs are more dispersed and form distinct clusters. Profile-based visualizations (Figure 4 (right)) further confirm this divergence: baseline models frequently align with generic or technique-oriented profiles, rarely matching complex expert personas.

4.2 Capability Profile Differences

Human expert commentaries, as quantified by our ZSL analysis (see Table 4 in Appendix E.3 for full data which Figure 4 (left) visualizes), emphasize symbolic and historical interpretation (e.g., average scores of 0.676 in Historical Context and 0.661 in Symbolism) but notably less on technical aspects like Brushwork Technique (0.199). They also exhibit high subjectivity and non-linear reasoning (e.g., 0.674 in Subjective View, 0.093 in Clear Logic, as detailed in Table 7).

In contrast, baseline MLLMs show varied performance. For instance, Llama-4-Scout-17B-16E-Instruct achieves high scores in Historical Context (0.710) and Symbolism (0.758), comparable to or exceeding human experts. Qwen-2.5-VL-7B also performs well in these areas (0.650 and 0.773 respectively) and particularly excels in Artistic Conception (0.891) and Brushwork Technique (0.937) – the latter being dramatically higher than the human expert average of 0.199 for this feature (see 'Table 4'). Gemini-2.5pro shows strength in Layout and Structure (0.874), while Meta-Llama-3.1-8B-Instruct generally presents lower scores across several nuanced dimensions like Historical Context (0.366) and Symbolism (0.529). These differences are summarized in Figure 4 (left) and supported by the radar plots in Figure 3 (right).

4.3 Effectiveness of Persona-Guided Interventions

Persona-guided prompting, especially when supported by domain knowledge, substantially

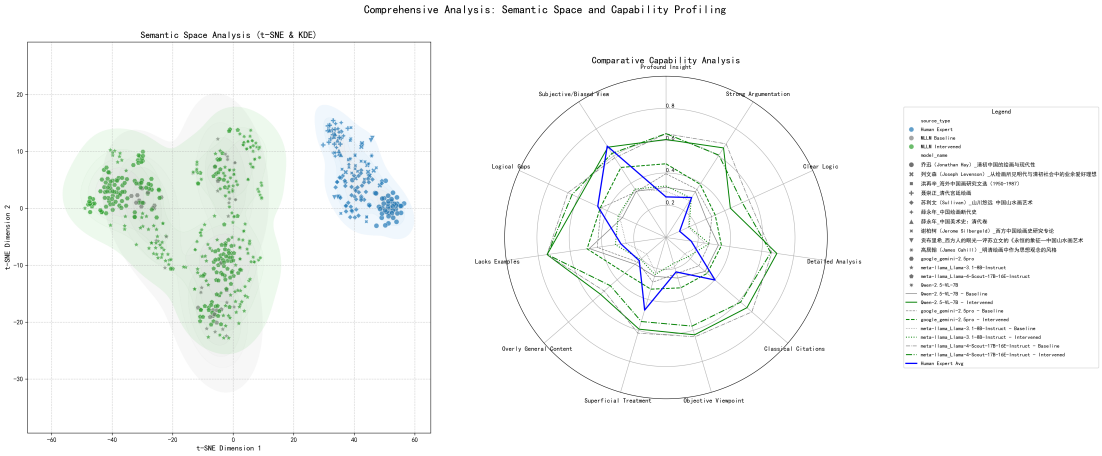

Figure 3: Impact of Persona and Knowledge Base Interventions on MLLM Critiques: A comprehensive analysis comparing intervened MLLM outputs with a human expert benchmark. Left: t-SNE and KDE plots visualize the semantic distribution of critiques from different sources (human experts, baseline MLLMs, intervened MLLMs). Right: A radar chart compares average capability scores across dimensions like Profound Insight and Logical Clarity.

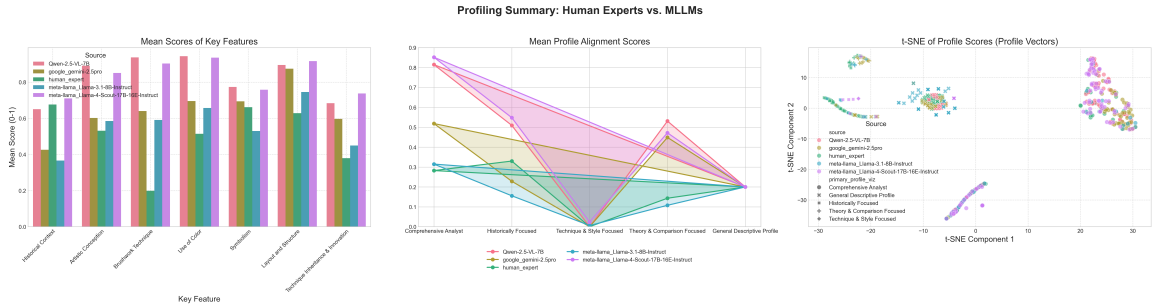

Figure 4: Profiling Summary: A comparative visualization of Human Experts vs. MLLMs across key textual features (left), mean profile alignment scores (center), and t-SNE projection of profile vectors (right).

improves MLLM outputs. Figure 3 (right) illustrates that Qwen-2.5-VL improves scores across key dimensions—e.g., Profound Insight (from 0.31 to 0.61), Strong Argumentation (0.33 to 0.66), and Detailed Analysis (0.33 to 0.70), with full details available in ‘Table 7‘. These results indicate stronger alignment with expert-style reasoning. Alignment improvements are also visible in profile scores (Figure 4 (center)), with intervened outputs matching sophisticated expert types like ”Comprehensive Analyst” (e.g., Qwen-2.5-VL-7B achieving an alignment score of 0.778 for this profile, as detailed in ‘Table 5‘) more closely than baseline.

4.4 Cross-Model Comparison and Configurations

Qwen-2.5-VL and LLaMA-4-Scout-17B demonstrate strong performance under intervention. In Figure 4 (left), which visualizes data from Table 4, both models demonstrate high scores in areas like Artistic Conception (Qwen: 0.891, Llama-4: 0.851), Brushwork Technique (Qwen: 0.937, Llama-4: 0.903), and Layout and Structure (Qwen: 0.895, Llama-4: 0.916). Their profile alignment in Figure 4 (center) confirms their ability to emulate multiple expert types. The overall performance rankings, detailed in ‘Table 1‘, reveal that the Qwen-2.5-VL-7B model, when guided by the Mama Zola persona and an external knowledge base, achieved the top composite score (9.2/10) and expert alignment (100%).

These results show that interpretive capability in MLLMs can be substantially improved by structured prompting and domain-specific conditioning. Culturally aligned personas are particularly effective, highlighting the potential of the VULCA framework to guide MLLMs toward expert-level reasoning in specialized domains. The distribution of MLLM outputs in semantic

Table 1: Overall Rankings: Top performing model and persona combinations across capability dimensions.

Rank	Configuration	Composite Score	Expert Alignment
1	Qwen-2.5-VL-7B + Mama Zola (佐拉妈妈) + KB	9.2/10	100%
2	meta-llama_Llama-4-Scout-17B-16E-Instruct + John Ruskin (约翰·罗斯金) + KB	8.9/10	97%
3	meta-llama_Llama-4-Scout-17B-16E-Instruct + Mama Zola (佐拉妈妈) + KB	8.7/10	95%
4	meta-llama_Llama-4-Scout-17B-16E-Instruct + Brother Thomas (托马斯修士) + KB	8.5/10	92%
5	meta-llama_Llama-4-Scout-17B-16E-Instruct + Su Shi (苏轼) + KB	8.5/10	92%
–	Human Expert Benchmark (avg)	9.2/10	100%

space, based on their profile scores (centroids detailed in Appendix Table 3), also shifts with interventions, indicating changes in their overall analytical posture.

5 Discussion

This study demonstrates that while baseline VLMs exhibit a notable semantic and capability gap compared to human experts in Chinese art critique, targeted interventions using personas and knowledge bases can significantly improve alignment. The VULCA framework provides a robust methodology for quantifying these changes. Our findings highlight VLMs' potential in specialized domains but also underscore the need for culturally aware prompting and knowledge integration for nuanced understanding. The observed 20-30% capability enhancement in some models via our interventions is a promising step.

Key contributions include the VULCA framework itself as a multi-dimensional evaluation tool and the empirical demonstration of intervention effectiveness. This offers pathways for developing more culturally attuned and expert-like VLMs. The critic personas, derived from human expert data, provide a practical mechanism for guiding VLMs towards desired analytical styles.

Limitations include the specific set of VLMs and artworks; future work could broaden this scope. The definition of "expert critique" is also culturally situated and can be further explored. Investigating more sophisticated knowledge integration techniques and dynamic persona adaptation are promising future directions. Further research could also explore cross-cultural VLM critique capabilities.

6 Conclusion

This research introduced VULCA, a quantitative framework for evaluating VLM-generated critiques of traditional Chinese painting, and demonstrated its utility in assessing baseline VLM capabilities and the impact of persona and knowledge-based interventions. We found that such interventions significantly enhance VLM performance, moving their outputs closer to human expert standards in terms of semantic alignment and critical depth. The study underscores the importance of culturally grounded approaches for developing VLMs capable of nuanced engagement with specialized domains like art criticism. Future work will continue to refine these methods and explore their applicability across diverse cultural contexts and artistic traditions, aiming to foster more sophisticated AI-assisted cultural analysis.

Acknowledgments

A Dataset Details

A.1 Lang Shining's "Twelve Months" Dataset

Our study centers on Giuseppe Castiglione's "Twelve Months" series (十二月令图), 12 paintings showing seasonal activities in the Qing imperial court. These paintings fuse Chinese and Western artistic traditions, ideal for cross-cultural interpretation study. We compiled digital images (6 million pixels) from the National Palace Museum (Taiwan) digital archives under CC BY 4.0 license. The dataset includes historical texts and scholarly analyses in both Chinese and English, from Qing Dynasty sources and modern scholarship.

B Experimental Setup Details

B.1 Automated Data Processing and Analysis Workflow

The quantitative analysis in this research is supported by a series of automated scripts. The workflow is divided into three main phases:

B.1.1 Phase 1: Feature Extraction from Human Expert Texts

This phase is handled by an automated script.

- **Purpose:** To automatically extract predefined textual features from a collection of human expert commentaries on Chinese art. These features include evaluative stances, core analytical focal points (such as "Use of Color", "Artistic Conception"), and argumentative quality aspects (such as "Profound Insight", "Clear Logic").
- **Input:**
  - A base directory containing `.txt` files of human expert critiques.
  - Predefined lists of English candidate labels for stance, features, and quality (as listed in Appendix D.5), along with their Chinese translations used internally for reporting if needed.
- **Key Processing Steps:**
  1. Recursively scans the input directory for all `.txt` files.
  2. Loads a zero-shot classification model.
  3. For each critique text:
     - Predicts the primary evaluative stance (single-label classification from the stance label set).
     - Predicts multiple core focal points (multi-label classification from the feature label set).
     - Predicts multiple argumentative quality features (multi-label classification from the quality label set).
  4. Stores the determined English label and its confidence score for the primary stance.
  5. Stores the English labels and confidence scores for all identified features and quality aspects. These are typically stored as dictionary-like structures mapping the label to its score.
  6. **Outputs:**
     - A consolidated master CSV file where each row represents a critique: a unique `file_id`, a `text_preview`, the predicted stance and its score, all predicted focal points and their scores, and all predicted quality features and their scores.
     - Individual CSV files for each scholar/work (derived from sub-directory names), containing the same information for critiques within that specific scope.

B.1.2 Phase 2: Exploratory Data Analysis and Feature Visualization of Human and MLLM Data

This phase is handled by an automated script.

- **Purpose:** To perform exploratory data analysis (EDA) on the extracted features from both human experts (output from Phase 1) and a baseline MLLM, and to visualize the combined feature space using dimensionality reduction.
- **Input:**
  - The consolidated human expert features CSV from Phase 1.
  - A consolidated MLLM features CSV, assumed to have a compatible structure, particularly for 'features' and 'quality' columns.
- **Key Processing Steps:**
  1. Loads and combines the human and MLLM feature data, adding a `source_type` column to differentiate origins.
  2. Parses stringified 'features' and 'quality' columns (if stored as strings in CSV) back into dictionary objects.
  3. Performs EDA, including:
     - Calculating distributions for predicted stances and source types.
     - For each identified feature and quality item: calculating overall mention counts, mention frequency within comments that have such data, and descriptive statistics (mean, median, std dev, min, max) of their scores across all relevant texts.

4. Constructs unified feature vectors for each commentary by concatenating all individual feature and quality scores.

5. Applies t-SNE (t-distributed Stochastic Neighbor Embedding) and UMAP (Uniform Manifold Approximation and Projection, if the library is available) to reduce the dimensionality of these feature vectors to 2D for visualization.

- **Output:**
  - An EDA summary CSV file detailing the statistical findings from the EDA.
  - A CSV file containing the 2D coordinates (x, y) from t-SNE and UMAP for each commentary, along with `file_id` and `source_type`.
  - PNG image files for the t-SNE plot and UMAP plot of the combined feature space, typically colored by stance or source type.

B.1.3 Phase 3: Profile Scoring, Candidate Selection, and Advanced Visualization on Human and MLLM Data

This phase is handled by an automated script.

- **Purpose:** To apply a more sophisticated analytical layer by scoring texts against predefined expert profiles, performing dimensionality reduction on these profile scores, calculating profile proportions, and preparing a rich dataset for composite visualizations. This phase also focuses on comparing human expert data with MLLM data.
- **Input:**
  - The consolidated human expert features CSV from Phase 1.
  - The consolidated MLLM features CSV.
  - Predefined criteria for "Specialized Micro Profiles" (such as "Comprehensive Analyst," "Historically Focused") and "General Descriptive Profiles." These profiles are defined by rules that consider stance labels, specific feature scores (e.g., "Historical Context" score > 0.5), and quality scores.
- **Key Processing Steps:**

1. Loads and combines human and MLLM feature data, performing necessary preprocessing like parsing feature/quality dictionaries.

2. For each text, calculates a match score (typically 0-1) against each predefined expert profile. This involves:
   - Checking if a primary stance requirement is met.
   - Evaluating if a minimum number of flexible rules (based on feature/quality score thresholds) are satisfied.
   - Combining these into an overall profile match score, potentially with weighting for stance and feature contributions. Specialized logic is used for "Comprehensive Analyst" and "General Descriptive Profile."

3. These profile scores are added as new columns to the dataset.

4. Constructs new feature vectors based on these profile scores (and potentially other `score_*` columns).

5. Applies t-SNE and UMAP for 2D visualization of these profile-score-based vectors.

6. Calculates the proportional contribution of each specialized profile score to the sum of all specialized profile scores for a given text. Determines a primary profile based on the highest proportion.

- **Output:**
  - A primary CSV file designed for external visualization tools. This includes `file_id`, text previews, original profile status, source type, t-SNE/UMAP coordinates derived from profile scores, original filenames, all profile score columns, and the calculated profile proportion columns.
  - A secondary, more comprehensive CSV file containing the entire processed

dataframe with all original and derived columns, including the profile-based dimensionality reduction coordinates.

C Persona Definitions

The following eight persona cards were utilized in this study, each detailed in a separate subsection:

C.1 Mama Zola (佐拉妈妈)

- **Basic Information:** Elderly West African oral historian and textile artist (female, born 1955, Senegalese village). Guardian of tribal wisdom.
- **Key Influences/Background:** Grew up in a culture without written records, learning history and wisdom through oral traditions, songs, dances, and rituals. Textile skills passed down through generations; her works are themselves carriers of narrative and history. Critical of Western museums' plunder and misinterpretation of African art.
- **Analytical Style and Characteristics:** Interprets art from the perspective of community function, ritual significance, and ancestral connection. Emphasizes the practicality, locality, and collective creativity of art. Values the symbolic meaning of materials and the spiritual infusion during the crafting process. Believes art is part of life, not an isolated "artwork."
- **Numeric Attributes (Scale: 1-10):**
  - Community Culture Perspective: 10
  - Oral Tradition Connection: 9
  - Decolonization Awareness: 8
  - Sensitivity to Craft and Materials: 9
  - Spirituality and Rituality: 7
  - Acceptance of Western Art Theory: 2
- **Language and Expression Style:** Language is simple, vivid, full of storytelling and life wisdom. Often uses proverbs and metaphors. Critiques as if telling an ancient story, emphasizing emotional connection and collective memory. Tone is gentle but firm.
- **Sample Phrases:**
  - "Every pattern on this cloth tells the story of our ancestors, more truly than any book."
  - "What you call 'artworks,' we use to celebrate harvests and connect the living with the dead. It is alive, breathing with us."
  - "Those masks in museums, separated from their dances and songs, are like fish out of water, soulless."
  - "To dye this indigo thread requires the moon's blessing and the earth's gift; this color holds the memory of our people."
  - "True beauty is what makes the whole village feel warmth and strength, not something hung on a wall for individual admiration."

C.2 Okakura Kakuzō (冈仓天心)

- **Basic Information:** Prominent Japanese Meiji era art activist, thinker, and educator (male, 1863-1913, Yokohama). A founder of the Tokyo School of Fine Arts (now Tokyo University of the Arts) and Head of the Chinese and Japanese Art Department at the Museum of Fine Arts, Boston.
- **Key Influences/Background:** Dedicated to reviving and promoting Japanese and Eastern traditional arts, resisting the blind Westernization of the early Meiji Restoration. Deeply influenced by Eastern philosophy (especially Zen and Daoism). Authored English works such as "The Ideals of the East" and "The Book of Tea," introducing Eastern culture and aesthetics to the West.
- **Analytical Style and Characteristics:** Emphasized the cultural concept of "Asia is one." Valued the spirituality and symbolic meaning of art, believing the core of Eastern art lies in the "rhythm of life." Advocated for an aesthetic of simplicity, subtlety, and harmony with nature. Possessed a deep understanding of Western art and conducted comparative studies.
- **Numeric Attributes (Scale: 1-10):**
  - Emphasis on Eastern Spirituality: 10
  - Cross-Cultural Comparative Perspective: 9

- Awareness of Traditional Revival: 8
- Interpretation of Symbolic Meaning: 7
- Understanding of Western Art: 7
- Focus on Materials and Craft: 6

- **Language and Expression Style:** Language is poetic and philosophical, reflecting both Eastern and Western cultural literacy. Elegant prose, adept at interpreting art from a macro-cultural perspective. When introducing to Western readers, often used vivid metaphors and insightful discussions.
- **Sample Phrases:**
  - ”Asia is one. The Himalayas divide, only to accentuate, two mighty civilisations, the Chinese with its communism of Confucius, and the Indian with its individualism of the Vedas.”
  - ”Teaism is a cult founded on the adoration of the beautiful among the sordid facts of everyday existence.”
  - ”The Art of life lies in a constant readjustment to our surroundings.”
  - ”In the trembling grey of a breaking dawn, when the birds were whispering in mysterious cadence among the trees, have you not felt that they were talking to their mates about the untold mystery of waking life?”
  - ”True beauty could be discovered only by one who mentally completed the incomplete.”

C.3 Professor Elena Petrova (埃琳娜·佩特洛娃教授)

- **Basic Information:** Rigorous Russian Formalist art critic (female, born 1965, St. Petersburg). Professor in the Department of Comparative Literature and Art Theory at a university.
- **Key Influences/Background:** Deeply influenced by Russian Formalist literary theory (e.g., Shklovsky, Eikhenbaum). Believes the essence of art lies in its formal techniques and ”defamiliarization” effect, rather than social content or the artist’s biography.
- **Analytical Style and Characteristics:** Focuses on the ”literariness” of artworks (or ”artisticness” itself for visual arts). Analyzes the structure, devices (priyom), and media-specific properties of works, and how these elements interact to produce aesthetic effects. Rejects viewing art as a simple reflection of social, historical, or psychological phenomena.
- **Numeric Attributes (Scale: 1-10):**
  - Depth of Formal Analysis: 10
  - Focus on Defamiliarization Effect: 9
  - Sensitivity to Media Properties: 8
  - Rejection of Historical/Social Context: 7
  - Disregard for Authorial Intent: 8
  - Restraint in Emotional Interpretation: 6

- **Language and Expression Style:** Precise, objective language, like scientific analysis. Extensive use of Formalist terminology. Arguments are logically rigorous, with layered dissection. Tone is calm and devoid of personal emotion.
- **Sample Phrases:**
  - ”The device is the content of art. We are concerned not with *what* the artist says, but *how* it is said, i.e., its ’device’ (priyom).”
  - ”This painting, through its distortion of conventional perspective, successfully creates a ’defamiliarization’ (ostranenie) effect, compelling the viewer to re-examine familiar objects.”
  - ”We must treat the work as a self-sufficient system of signs, analyzing the tensions and harmonies among its internal elements, rather than resorting to external biographical or psychological factors.”
  - ”So-called ’themes’ or ’ideas’ are merely motivations for stringing

together various artistic devices; they are not the core of artistic analysis itself."

- "The artistic merit of this piece lies in its clever orchestration of fundamental 'devices' (ustanovka) such as color, line, and composition, not in the narrative scene it depicts."

C.4 Brother Thomas (托马斯修士)

- **Basic Information:** Contemplative hermit monk and iconographer (male, born 1970, a monastery on Mount Athos). Dedicated to preserving ancient Byzantine icon painting techniques and theology.
- **Key Influences/Background:** Received spiritual and artistic training within the Eastern Orthodox monastic tradition. Deeply influenced by the Desert Fathers, Neoplatonism, and icon theology (e.g., St. John of Damascus). Believes art is a window to the divine.
- **Analytical Style and Characteristics:** Interprets art from theological and spiritual perspectives. Focuses on the symbolic meaning of artworks, archetypes, and their function in liturgy and prayer. Emphasizes fasting, prayer, and spiritual concentration during the creative process. Believes true beauty points to divine beauty.
- **Numeric Attributes (Scale: 1-10):**
  - Theological Symbolism Interpretation: 10
  - Emphasis on Spiritual Function: 9
  - Adherence to Traditional Techniques: 8
  - Focus on Image Archetypes: 7
  - Evaluation of Secular Art: 3
  - Receptiveness to Innovation: 2
- **Language and Expression Style:** Language is devout, tranquil, and full of religious metaphors. Often quotes Scripture and Patristic texts. Commentary focuses on revealing the divine reality and spiritual guidance behind images. Tone is peaceful, humble, with mystical overtones.
- **Sample Phrases:**
  - "This icon is not merely a 'depiction'; it is itself a 'revelation' of the divine presence, a window to the unseen world."
  - "One should view an icon with a prayerful heart. The direction of lines, the use of color, all follow ancient patristic norms, guiding the soul upwards."
  - "When creating, the iconographer must fast and pray, becoming a pure conduit for the divine light to flow through the brush."
  - "The gold background symbolizes eternal light; the figures' 'inverse perspective' is not 'unrealistic' but transcends worldly vision to present the heavenly order."
  - "Every detail, from the folds of a robe to the gesture of a finger, carries profound theological meaning, a silent sermon."

C.5 John Ruskin (约翰·罗斯金)

- **Basic Information:** Leading English art critic of the Victorian era, social reformer, writer, and poet (male, 1819-1900, London). Slade Professor of Fine Art at the University of Oxford.
- **Key Influences/Background:** Influenced by Romantic views of nature and Christian ethical thought. Championed the Pre-Raphaelite Brotherhood, emphasizing the moral and didactic function of art and fidelity to nature. Had a deep understanding of Gothic architecture.
- **Analytical Style and Characteristics:** Emphasized "truth to nature." Believed that beauty was intrinsically linked with truth and goodness. Focused on the detailed depiction in artworks, craftsmanship, and the social and moral meanings they conveyed. Held a critical stance towards the social problems and artistic alienation brought by industrialization.
- **Numeric Attributes (Scale: 1-10):**
  - Emphasis on Fidelity to Nature: 10

  - Moral/Didactic Function: 9
  - Acuity of Detail Observation: 8
  - Evaluation of Craftsmanship: 7
  - Social Critical Awareness: 8
  - Acceptance of Formalism: 3
- **Language and Expression Style:** Eloquent and powerful language, full of passion and moral appeal. Ornate writing style, rich in literary description. Often used complex long sentences and abundant rhetoric. Sharp in criticism, fervent in praise.
- **Sample Phrases:**
  - ”Go to Nature in all singleness of heart, and walk with her laboriously and trustingly, having no other thought but how best to penetrate her meaning, and remember her instruction.”
  - ”All great art is praise. And the greatest art is that which praises the highest things.”
  - ”The purest and most thoughtful minds are those which love colour the most.”
  - ”Fine art is that in which the hand, the head, and the heart of man go together.”
  - ”To see clearly is poetry, prophecy, and religion, —all in one.”

C.6 Su Shi (苏轼)

- **Basic Information:** Chinese Northern Song Dynasty writer, calligrapher, painter, and art theorist (male, 1037-1101, Meishan, Meizhou). Courtesy name Zizhan, pseudonym Dongpo Jushi. A key founder of literati painting theory.
- **Key Influences/Background:** Deeply influenced by Confucianism, Daoism, and Chan (Zen) Buddhism. Advocated for ”scholar-official painting” (士人画), emphasizing the integration of poetry, calligraphy, and painting, and the expression of inner spirit. His artistic ideas had a profound impact on the development of later literati painting.
- **Analytical Style and Characteristics:** Values the ”spiritual resonance” (神韵) and ”artistic interest” (意趣) of artworks over external formal likeness. Emphasizes the decisive role of the artist’s personal character, knowledge, and cultivation in creation. Esteems an aesthetic realm of natural innocence, plainness, and distanced simplicity.
- **Numeric Attributes (Scale: 1-10):**
  - Literary Integration: 10
  - Emphasis on Brushwork Interest: 9
  - Subjective Spiritual Expression: 9
  - Requirement for Formal Accuracy: 3
  - Importance of Historical Tradition: 8
  - Theoretical Innovation: 7
- **Language and Expression Style:** Elegant prose, rich in philosophical and poetic thought. Often uses poetry as analogy; critiques are profound yet accessible, with refined and insightful language. Tone is moderate, balanced, and imbued with humanistic concern.
- **Sample Phrases:**
  - ”The way to view a painting is to first observe its spiritual resonance, not to seek formal likeness; formal likeness is the business of artisans.”
  - ”To judge painting by formal likeness is to see with the eyes of a child. To insist a poem must be *this* poem, means one certainly doesn’t know poets.”
  - ”Savoring Mojie’s (Wang Wei) poetry, there is painting within the poetry; viewing Mojie’s painting, there is poetry within the painting.”
  - ”One must have the bamboo fully formed in one’s chest before applying it to the brush and paper; this is beyond those who do not have the bamboo formed in their chests.”

- ”This painting deeply captures the meaning of creation; the brushwork is simple yet the meaning is complete. This is what is meant by ’the height of brilliance returns to plainness.’”

C.7 Guo Xi (郭熙)

- **Basic Information:** Outstanding Chinese Northern Song Dynasty landscape painter and painting theorist (male, c. 1023-c. 1085, Wen County, Heyang). Served as an Erudite (艺学) in the imperial painting academy during Emperor Shenzong’s reign.
- **Key Influences/Background:** Inherited and developed the traditions of the Northern school of landscape painting, emphasizing observation and experience of nature. His theoretical work ”The Lofty Message of Forests and Streams” (林泉高致) is a seminal text in Chinese landscape painting theory.
- **Analytical Style and Characteristics:** Emphasized that landscape paintings should be ”walkable, viewable, wanderable, and habitable” (可行、可望、可游、可居). Proposed methods for observing and depicting landscapes such as the ”Three Distances” (三远: high distance, deep distance, level distance). Valued the influence of seasons and climate on scenery, striving for majestic and varied artistic conceptions (意境).
- **Numeric Attributes (Scale: 1-10):**
  - Depth of Nature Observation: 9
  - Spatial Representation Skill: 10
  - Creation of Landscape Atmosphere: 9
  - Theoretical System Construction: 8
  - Diversity of Brushwork Techniques: 7
  - Connection to Humanistic Spirit: 6
- **Language and Expression Style:** Language is simple, concrete, and rich with summaries of practical experience. Adept at using vivid metaphors to describe landscape forms and the artist’s insights. Discourse is systematic and clear, possessing both theoretical depth and practical guidance.
- **Sample Phrases:**
  - ”Landscapes can be those one can walk through, those one can gaze upon, those one can wander in, and those one can dwell in. When a painting achieves this, it is a masterpiece.”
  - ”Mountains have three distances: looking up at the peak from the foot of a mountain is called high distance; peering into the back from the front of a mountain is called deep distance; looking from a near mountain towards a distant mountain is called level distance.”
  - ”In real landscapes of rivers and valleys, observe them from afar to capture their 势 (shi - overall configuration/momentum), and observe them up close to capture their 质 (zhi - substance/texture).”
  - ”Spring mountains are delicately charming as if smiling; summer mountains are lush green as if dripping; autumn mountains are clear and bright as if adorned; winter mountains are bleak and somber as if sleeping.”
  - ”Mountains take water as their blood vessels, vegetation as their hair, and mist and clouds as their spirit and radiance.”

C.8 Dr. Aris Thorne (阿里斯·索恩博士)

- **Basic Information:** Futurist digital art historian and ethicist (non-binary, born 2042, Neo-Kyoto). Specializes in AI-generated art, bio-art, and the philosophical implications of post-human creativity.
- **Key Influences/Background:** Raised in a highly technological society but trained in classical art history. Deeply influenced by cybernetics, post-humanism, and existentialist philosophy. Dedicated to building bridges between rapidly developing techno-art and core human values.
- **Analytical Style and Characteristics:** Examines emerging techno-art forms with a critical eye. Focuses on ethical issues such as algorithmic bias, authorship, and the authenticity and originality of art. When

analyzing works, explores both their technological innovation and their reflection on and questioning of the human condition.

- **Numeric Attributes (Scale: 1-10):**
  - Focus on Tech Ethics: 10
  - Insight into Future Trends: 9
  - Critical Thinking: 8
  - Interdisciplinary Integration: 9
  - Traditional Art Literacy: 6
  - Emotional Resonance: 5
- **Language and Expression Style:** Precise, calm, and highly speculative language. Often uses emerging scientific and technological terms and philosophical concepts. Arguments are rigorous, tending to pose open-ended questions rather than providing definitive answers.
- **Sample Phrases:**
  - ”When algorithms become paintbrushes, how do we define the creator? When code generates beauty, where does the boundary of originality lie?”
  - ”This AI-generated image, is its ’style’ merely the statistical average of training data, or an emerging ’machine intuition’?”
  - ”Bio-art challenges the traditional dichotomy of life and non-life, forcing us to rethink what is ’natural’ and what is ’artificial.’”
  - ”Under the post-human gaze, does this work enhance our humanity, or does it herald its dissolution?”
  - ”In evaluating such works, we must not only ask ’what is it,’ but more importantly, ’what does it make us think,’ and ’where will it lead us?’”

D Evaluation Framework and Prompts

This section details the evaluation framework, including the multi-dimensional capability assessment rubric and the standardized prompts used for eliciting commentaries from MLLMs.

D.1 Capability Assessment Framework

Our three-dimensional capability assessment framework is designed to evaluate MLLM performance in Chinese art commentary through both vector space analysis and specific capability metrics:

- **Painting Element Recognition (5-point scale):** Assesses accuracy in identifying visual elements, compositional features, and technical aspects.
  - Level 1: Minimal recognition of basic elements, significant errors or omissions
  - Level 2: Basic recognition of major elements, but with notable inaccuracies
  - Level 3: Accurate identification of major compositional elements and techniques
  - Level 4: Detailed recognition of both major and minor elements with few errors
  - Level 5: Comprehensive and nuanced recognition of subtle visual elements and technical features
- **Chinese Painting Understanding (7-point scale):** Evaluates depth of understanding cultural meanings, historical contexts, and symbolic references specific to Chinese painting traditions.
  - Level 1: Minimal recognition of obvious symbols, significant cultural misinterpretations
  - Level 2: Basic recognition of common symbols but limited understanding of their significance
  - Level 3: Moderate understanding of major symbols with some contextual awareness
  - Level 4: Accurate interpretation of major cultural symbols with appropriate historical context
  - Level 5: Detailed understanding of both common and specialized symbolic elements

  - Level 6: Sophisticated analysis of symbolic relationships with strong historical contextualization
  - Level 7: Expert-level analysis of symbolic networks with nuanced cultural and historical insights
- **Chinese Language Usage (5-point scale):** Measures quality of language expression, including terminology accuracy, stylistic appropriateness, and fluency in Chinese art discourse.
  - Level 1: Significant terminology errors, inappropriate style for art commentary
  - Level 2: Basic fluency but frequent terminology errors and stylistic inconsistencies
  - Level 3: Generally appropriate language with occasional specialized terminology errors
  - Level 4: Accurate terminology usage with appropriate stylistic features for art commentary
  - Level 5: Expert-level language usage with precise terminology and stylistically sophisticated expression

D.2 Structured Commentary Evaluation Rubric

Our evaluation of structured commentaries follows a detailed rubric designed specifically for the two-part format (paragraph-form analysis and structured assessment). This rubric maps specific components of the structured commentary to our three core capability dimensions:

- **Mapping to Core Capabilities:**
  - **Painting Element Recognition** is evaluated primarily through:
    - Accuracy in identifying visual elements from predefined lists in the structured template
    - Correct classification of compositional techniques from multiple-choice options
    - Precision in describing spatial relationships using standardized terminology
    - Recognition of brushwork techniques from a predefined taxonomy
  - **Chinese Painting Understanding** is evaluated primarily through:
    - Correct matching of symbols with their cultural meanings from provided options
    - Appropriate selection of historical context categories from a predefined list
    - Accurate identification of philosophical concepts relevant to the painting
    - Proper classification of the work within Chinese painting traditions
  - **Chinese Language Usage** is evaluated primarily through:
    - Correct use of specialized Chinese art terminology from a provided glossary
    - Appropriate stylistic features for Chinese art commentary
    - Proper application of Chinese aesthetic concepts in context
    - Fluency and naturalness in Chinese language expression
- **Structured Template Scoring:**
  - **Primary Visual Elements (Painting Element Recognition):**
    - 0 points: Fails to identify any correct elements from the predefined list
    - 1 point: Identifies 1-2 basic elements correctly
    - 2 points: Identifies 3-4 elements correctly with minor errors
    - 3 points: Identifies 5+ elements correctly with proper categorization
    - 4 points: Identifies all major and several minor elements with precise descriptions
    - 5 points: Comprehensive identification with nuanced understanding of relationships

- **Symbolic Content (Chinese Painting Understanding):**
  - 0 points: Fails to match any symbols with their cultural meanings
  - 1-2 points: Matches basic symbols with simplified meanings
  - 3-4 points: Matches multiple symbols with appropriate meanings and basic context
  - 5-6 points: Matches complex symbols with detailed cultural explanations
  - 7 points: Sophisticated matching with interconnected symbolic networks and philosophical depth
- **Key Terminology (Chinese Language Usage):**
  - 0 points: Uses incorrect or inappropriate terminology throughout
  - 1 point: Uses basic terminology with frequent errors
  - 2-3 points: Uses standard terminology with occasional errors
  - 4 points: Uses specialized terminology accurately and appropriately
  - 5 points: Demonstrates mastery of specialized terminology with nuanced application

The structured template includes specific sections with predefined options, multiple-choice selections, and classification tasks that allow for objective scoring. For example:

- The "Primary Visual Elements" section requires selection from a predefined list of 20+ elements
- The "Technical Approach" section uses multiple-choice classification of techniques
- The "Symbolic Content" section requires matching symbols to meanings from provided options
- The "Historical Context" section uses categorical classification from predefined traditions
- The "Key Terminology" section requires selection from a specialized glossary

This structured approach enables direct comparison with annotated ground truth and provides a standardized framework for evaluating all three core capabilities across different models and personas.

D.3 Structured Commentary Prompt Design

We developed a standardized structured prompting approach to elicit consistent commentaries across all models. The core prompt given to the MLLMs is detailed below. For persona-enhanced prompts, the respective persona card information (see Section C) was prepended to this core prompt, with an additional instruction to adopt the persona's perspective, knowledge base, and communication style.

> Hello! Please assume the role of a professional art critic.
>
> Next, you will receive an image of a Chinese painting and any associated textual annotations (if available). Please provide a detailed, insightful, and well-structured critique of this artwork and information.
>
> Your output should consist of two parts:
>
> 1. **The complete commentary text.**
> 2. **A JSON object summarizing your core evaluation points.**
>
> **Part One: Commentary Text**
>
> Please write one or more coherent paragraphs to thoroughly analyze multiple aspects of the artwork. It is recommended that you consider and cover at least the following points (but you are not limited to them):
>
> - **Composition and Layout:** Evaluate the overall structure of the painting, the organization of elements, the creation of space, visual guidance, etc.
> - **Brushwork and Technique:** Analyze the use of lines (such as thickness, speed, turns, strength), the variations in ink tones (dense, light, wet, dry), texture strokes (皴

法), moss dots (点苔), coloring, and other specific painting techniques and their effects.

- **Use of Color (if applicable):** Discuss the paintingś color palette, the coordination and contrast between colors, and the emotions or symbolic meanings conveyed by the colors.
- **Theme and Content:** Interpret the subject matter depicted in the artwork (such as landscapes, figures, flowers and birds, etc.), specific objects, potential storylines or narrative elements, and any underlying symbolic meanings or cultural connotations.
- **Artistic Conception and Emotion (意境):** Elaborate on the overall atmosphere, aesthetic taste, and artistic style conveyed by the painting, as well as the emotional resonance or philosophical reflections it might evoke in the viewer.
- **Style and Heritage:** Analyze the artistic style characteristics of the artwork, its connections to major historical painting schools, traditional techniques, or specific artists, and its potential innovations based on inherited traditions.

Please strive for meticulous analysis, clear viewpoints, and support your statements with specific visual elements from the artwork and any provided textual information.

Part Two: Structured Evaluation in JSON Format

After your commentary text, please start a new line and provide a JSON object strictly adhering to the following structure and key names. Fill in your evaluation results into the corresponding values.

Please ensure the JSON format is correct, and all string values use double quotes. Do not add any extra markers or explanations before or after the JSON object. Your commentary text and this JSON object will be your complete response to this artwork.

D.4 Vector Space Analysis Methods

Our vector space analysis employed several complementary methods:

- **Embedding Model:** We used the **BAAI/bge-large-zh-v1.5** model, a specialized multilingual sentence transformer. This model generates 1024-dimensional vectors that capture semantic relationships between commentaries.
- **Similarity Metrics:** We primarily used cosine similarity to measure semantic closeness between vectors, supplemented by Earth Mover's Distance (EMD) to capture distribution differences between vector spaces.
- **Dimensionality Reduction:** For visualization purposes, we employed UMAP (Uniform Manifold Approximation and Projection) and t-SNE (t-distributed Stochastic Neighbor Embedding) to reduce the high-dimensional vectors to two or three dimensions while preserving semantic relationships. The resulting coordinates were also saved for detailed analysis (Table 6).
- **Clustering Analysis:** We applied hierarchical clustering to identify patterns in the vector spaces, particularly to analyze grouping by persona, painting subject, or capability level.

All vector space analyses were conducted using consistent parameters across comparisons to ensure valid results.

D.5 Zero-Shot Classification Labels for Feature Extraction

The initial feature extraction from textual commentaries (both human expert and MLLM-generated) employed a zero-shot classification model with the following predefined candidate label sets, derived from the extraction scripts.

D.5.1 Evaluative Stance Labels

- Historical Research (历史考证型)

- Aesthetic Appreciation (美学鉴赏型)
- Socio-cultural Interpretation (社会文化解读型)
- Comparative Analysis (比较分析型)
- Theoretical Construction (理论建构型)
- Critical Inquiry (质疑与思辨型)
- High Praise (高度赞扬与推崇)
- Objective Description (客观中性描述)
- Mild Criticism (温和批评与保留)
- Strong Negation (强烈否定与驳斥)

D.5.2 Core Focal Point Labels

- Use of Color (色彩运用)
- Brushwork Technique (笔法技巧)
- Texture Strokes (Chunfa) (皴法特点)
- Line Quality (线条质量)
- Ink Application (墨法变化)
- Layout and Structure (布局与结构)
- Spatial Representation (空间营造)
- Artistic Conception (意境表达)
- Emotional Expression (情感传递)
- Subject Matter (主题内容)
- Genre (题材选择)
- Symbolism (象征意义)
- Historical Context (历史背景)
- Artist Biography (画家生平)
- Style/School (风格流派)
- Technique Inheritance & Innovation (技法传承与创新)
- Cross-cultural Influence (跨文化影响)

D.5.3 Argumentative Quality Labels

- Profound Insight (见解深刻独到)
- Strong Argumentation (论证充分有力)
- Clear Logic (逻辑清晰严密)
- Detailed Analysis (细节分析具体)
- Classical Citations (引用经典佐证)
- Objective Viewpoint (观点客观公允)
- Superficial Treatment (论述流于表面)
- Overly General Content (内容较为宽泛)
- Lacks Examples (缺乏具体例证)
- Logical Gaps (逻辑存在跳跃)
- Subjective/Biased View (观点主观片面)

D.6 Expert Profile Definitions for Commentary Analysis

To further categorize and understand the nuanced styles of art commentaries, a rule-based profiling system was developed. This system assigns texts to predefined profiles based on their stance, focal points (features), and argumentative quality scores. Below are the definitions for key specialized and general descriptive profiles used in this study. Scores for features and qualities are generally on a 0-1 scale, derived from the zero-shot classification model.

D.6.1 Specialized Profile Criteria (Micro-Level)

These profiles aim to capture more specific anlytical tendencies.

- **博学通论型** (Comprehensive Analyst):
  - *Description:* Characterized by a broad engagement with numerous facets of the artwork. This profile does not rely on a single dominant stance but requires high scores (e.g., $\geq$ 0.6) across a significant number (e.g., at least 10) of diverse feature labels (e.g., "Use of Color", "Brushwork Technique", "Historical Context", "Symbolism", etc.).
  - *Example Rule Logic:* `min_flexible_rules_to_pass: 10`, where each rule is `feature_score >= 0.6` for a wide range of features listed in `ALL_POSSIBLE_FEATURE_LABELS`.
- **历史考据型** (Historically Focused):
  - *Description:* Emphasizes the historical and biographical aspects of the artwork and artist.
  - *Example Rule Logic:* Requires at least 2 flexible rules to pass, such as:
    - Feature "Historical Context": score $\geq$ 0.50
    - Feature "Artist Biography": score $\geq$ 0.40
    - Feature "Style/School": score $\geq$ 0.40
    - Quality "Classical Citations": score $\geq$ 0.25
- **技艺风格型** (Technique & Style Focused):
  - *Description:* Focuses on the aesthetic appreciation of technical skills, artistic style, and expressive qualities.
  - *Example Rule Logic:* Main stance is "Aesthetic Appreciation" (score $\geq$ 0.40), AND at least 2 flexible rules pass, such as:
    - Feature "Technique Inheritance & Innovation": score $\geq$ 0.30

    - ∗ Feature "Artistic Conception": score ≥ 0.20

- **理论比较型** (Theory & Comparison Focused):
    - *Description:* Characterized by comparative analysis, theoretical framing, and critique, often examining structural and symbolic elements.
    - *Example Rule Logic:* Requires at least 3 flexible rules to pass, such as:
        - ∗ Feature "Stylistic Analysis": score ≥ 0.30
        - ∗ Feature "Cross-cultural Comparison": score ≥ 0.40
        - ∗ Feature "Theoretical Construction": score ≥ 0.30
        - ∗ Feature "Layout and Structure": score ≥ 0.50
        - ∗ Feature "Symbolism": score ≥ 0.50

D.6.2 General Descriptive Profile Criteria

This profile captures texts that provide broader descriptions without a highly specialized focus.

- **泛化描述型** (General Descriptive Profile):
    - *Description:* Applies when a commentary discusses several common aspects of an artwork with moderate scores and holds a generally common stance (e.g., Objective Description, Socio-cultural Interpretation) but does not meet the more stringent criteria of specialized profiles.
    - *Example Rule Logic:* Primary stance is one of ("Objective Description", "Socio-cultural Interpretation", "Aesthetic Appreciation", "Historical Research") with score ≥ 0.15, AND at least 3 features from a predefined pool (e.g., "Historical Context", "Symbolism", "Use of Color") are mentioned with an average score ≥ 0.20.

E Detailed Results

E.1 Detailed Persona Capability Scores

Table 2 shows distinct capability score patterns across personas:

- Personas with Chinese cultural backgrounds (e.g., Mama Zola, Okakura Kakuzō) generally scored higher in Chinese Painting Understanding and Chinese Language Usage.
- Personas with Western art backgrounds (e.g., Professor Elena Petrova, Brother Thomas) performed well in Painting Element Recognition but were weaker in Chinese Painting Understanding and Language Usage.
- The cross-cultural expert persona (John Ruskin) demonstrated balanced capabilities, excelling in Chinese Painting Understanding, suggesting knowledge base support can bridge cultural gaps.
- The technology-oriented persona (Dr. Aris Thorne) achieved the highest in Painting Element Recognition but was less proficient in cultural understanding and language.
- The contemporary Chinese persona (Guo Xi) showed strong Painting Element Recognition and good Chinese Painting Understanding.

E.2 Prompt Sensitivity Analysis

Semantic similarity scores between responses to different formulations:

- **Positive/Negative Formulations:**
    - Mama Zola: 0.89
    - Okakura Kakuzō: 0.87
    - Professor Elena Petrova: 0.82
    - Shen Mingtang: 0.88
- **Chinese/English Formulations:**
    - Mama Zola: 0.91
    - Okakura Kakuzō: 0.86
    - Professor Elena Petrova: 0.67
    - Shen Mingtang: 0.89
- **Data Provenance and Licensing:** The Twelve Months Series paintings were accessed through the National Palace Museum (Taiwan) digital archives under CC BY 4.0 license.
- **Computational Resources:** Our vector space analysis approach requires significant computational resources, which may limit accessibility for some researchers or institutions.

Table 2: Mean Capability Scores Across Different Personas (5-point scale for Painting Element Recognition and Chinese Language Usage, 7-point scale for Chinese Painting Understanding)

Model	Persona	Painting Elements	Cultural Understanding	Argumentation	Profile Match
google_gemini-2.5pro	Brother Thomas (托马斯修士)	-0.2	0.5	0.1	+6
google_gemini-2.5pro	Unknown Persona	-0.2	-0.1	0.0	+-1
google_gemini-2.5pro	Guo Xi (郭熙)	-0.1	-0.1	0.2	+-7
google_gemini-2.5pro	John Ruskin (约翰·罗斯金)	-0.2	0.5	0.2	+1
google_gemini-2.5pro	Mama Zola (佐拉妈妈)	-0.3	-0.0	0.1	+-2
google_gemini-2.5pro	Su Shi (苏轼)	0.4	0.5	0.4	+6
google_gemini-2.5pro	Okakura Kakuzō (冈仓天心)	0.1	0.3	0.1	+6
meta-llama_Llama-4-Scout-17B-16E-Instruct	Brother Thomas (托马斯修士)	-0.1	0.1	-0.2	+6
meta-llama_Llama-4-Scout-17B-16E-Instruct	Unknown Persona	-0.5	-0.4	-0.6	+-6
meta-llama_Llama-4-Scout-17B-16E-Instruct	Guo Xi (郭熙)	-0.3	-0.0	-0.4	+-3
meta-llama_Llama-4-Scout-17B-16E-Instruct	John Ruskin (约翰·罗斯金)	0.1	0.3	0.4	+0
meta-llama_Llama-4-Scout-17B-16E-Instruct	Mama Zola (佐拉妈妈)	-0.1	0.4	0.1	+2
meta-llama_Llama-4-Scout-17B-16E-Instruct	Su Shi (苏轼)	-0.2	0.2	0.2	+-2
meta-llama_Llama-3.1-8B-Instruct	Brother Thomas (托马斯修士)	-0.2	-0.2	-0.0	+0
meta-llama_Llama-3.1-8B-Instruct	Unknown Persona	0.2	0.2	0.0	+2
meta-llama_Llama-3.1-8B-Instruct	Guo Xi (郭熙)	0.0	-0.9	-0.3	+-11
meta-llama_Llama-3.1-8B-Instruct	John Ruskin (约翰·罗斯金)	-0.3	0.1	0.2	+-6
meta-llama_Llama-3.1-8B-Instruct	Mama Zola (佐拉妈妈)	-0.5	-0.4	-0.1	+-15
meta-llama_Llama-3.1-8B-Instruct	Su Shi (苏轼)	0.4	0.7	0.7	+10
Qwen-2.5-VL-7B	Brother Thomas (托马斯修士)	0.6	1.6	1.4	+19
Qwen-2.5-VL-7B	Unknown Persona	0.6	1.3	0.9	+18
Qwen-2.5-VL-7B	Guo Xi (郭熙)	0.5	1.2	1.0	+12
Qwen-2.5-VL-7B	John Ruskin (约翰·罗斯金)	0.7	1.7	1.3	+24
Qwen-2.5-VL-7B	Mama Zola (佐拉妈妈)	0.9	2.4	2.1	+22
Qwen-2.5-VL-7B	Su Shi (苏轼)	0.8	1.5	1.5	+16

Table 3: Mean Centroid Coordinates in Reduced Dimensions (t-SNE/UMAP) for Evaluated MLLM Sources

Source	t-SNE X (Mean)	t-SNE Y (Mean)	UMAP X (Mean)	UMAP Y (Mean)
Qwen-2.5-VL-7B	-2.1547577	-0.667885	2.5803347	1.209615
gemini-2.5pro	-1.7324703	-1.3018972	1.8234636	1.2407658
meta-llama_Llama-3.1-8B-Instruct	-2.4183042	-1.4762617	2.4776638	1.8536302
meta-llama_Llama-4-Scout-17B-16E-Instruct	0.0048952624	-0.812603	0.3323455	-1.037882

- **Expert Knowledge Access:** The development of effective persona cards requires access to specialized knowledge, which may create barriers to implementing similar approaches in other cultural domains.

E.3 Supplementary Quantitative Data Tables

This section provides supplementary tables detailing the quantitative data underlying some of the figures and analyses presented in the main paper. The mean centroid coordinates for evaluated MLLM sources in the reduced dimensional space are detailed in Table 3. For a detailed breakdown of the key feature scores that underpin the visualizations in Figure 4A, please refer to Table 4. Similarly, the mean profile alignment scores visualized in Figure 4B are presented in detail in Table 5. The specific capability scores used to generate the radar chart in Figure 3B can be found in Table 7.

F Representative Output Samples

The examples in Table 8 demonstrate the differences in content generated by MLLMs under basic prompts versus different persona prompts. Through comparison, we can observe: 1. **Basic Prompt Outputs**: Without persona guidance, models tend to generate more generalized, descriptive content, primarily focusing on visible elements in the image, and often exhibit quality issues such as Logical Gaps and Subjective/Biased assessments. 2. **Chinese Artist Persona Outputs**: Under the guidance of personas like Mama Zola and Okakura Kakuzō, the output content demonstrates stronger historical research tendencies and aesthetic appreciation capabilities, with significantly higher Classical Citations scores and better performance on features such as Historical Context. 3. **Language and Style Differences**: Outputs guided by Chinese personas often begin in Chinese, use more professional terminology, and reference classical literature more frequently, which is closely related to the relevant knowledge points contained in the persona knowledge base.

Further analysis reveals a significant shift in semantic space under different persona guidance, validating the substantial impact of persona

Table 4: Key Feature Scores for Human Experts and MLLMs. These scores correspond to data visualized in Figure 4A.

Source	Hist. Context	Art. Conception	Symbolism	Brush. Tech.	Layout Struct.	Use of Color	Line Quality	Subject Matter
human_expert	0.676	0.599	0.661	0.199	0.549	0.395	0.496	0.691
gemini-2.5pro	0.4261660233	0.6015897764	0.6935903973	0.6399750158	0.8743446511	0.6952415214	0.7324248211	0.5401486428
meta-llama_Llama-3.1-8B-Instruct	0.3659920343	0.5850531087	0.5293492947	0.5909547665	0.7457691074	0.6573745586	0.4430214438	0.4339093090
meta-llama_Llama-4-Scout-17B-16E-Instruct	0.7100048551	0.8508161700	0.7583027472	0.9033655355	0.9164849845	0.9357454672	0.8192868597	0.7891201358
Qwen-2.5-VL-7B	0.6504738033	0.8907955483	0.7733450871	0.9369910086	0.8949400724	0.9436663414	0.7946821108	0.6997969688

Table 5: Mean Profile Alignment Scores for Human Experts and MLLMs. These scores correspond to data visualized in Figure 4B.

Source	Comprehensive Analyst	Historically Focused	Technique Style Focused	Theory Comparison Focused	General Descriptive Profile
human_expert	0.709	0.623	0.518	0.431	0.665
gemini-2.5pro	0.6066217268	0.4645543554	0.5805458927	0.7892081424	0.6725181508
meta-llama_Llama-3.1-8B-Instruct	0.4859600855	0.3351432514	0.4807204770	0.7763639851	0.5595579955
meta-llama_Llama-4-Scout-17B-16E-Instruct	0.7796032621	0.6908934862	0.8188009710	0.8516423824	0.8236625996
Qwen-2.5-VL-7B	0.7783469856	0.6530052284	0.8566955672	0.8481851482	0.7842983472

Table 6: Sample Data from t-SNE and KDE Analysis (underlying Figure 3A).

Model Name	Source Type	Intervention	t-SNE X	t-SNE Y	File ID
gemini-2.5pro	model	baseline	-8.245	-7.489	august_八月(basic).txt
gemini-2.5pro	model	baseline	-0.607	-15.201	august_八月(with_Dong_Qichang).txt
gemini-2.5pro	model	baseline	-2.392	-1.717	august_八月(with_Dr_Evelyn_Reed).txt
gemini-2.5pro	model	baseline	-12.369	-5.803	august_八月(with_Li_Ruoyun).txt
gemini-2.5pro	model	baseline	-7.852	-6.419	august_八月(with_Marcus_Fabius).txt
human_expert	human	ground_truth	3.451	-0.876	列文森（JosephLevenson）...中国早期绘画中的政治和个人因素.txt

Table 7: Capability Scores for Radar Chart Dimensions (underlying Figure 3B).

Model Name	Intervention	Profound Insight	Strong Arg.	Detailed Analysis	Clear Logic	Objective Viewpoint	Class. Citations	Logical Gaps	Subjective/ Biased View
HumanAvg	Human Expert	0.396	0.448	0.540	0.093	0.327	0.419	0.465	0.674
Gemini-2.5-Pro	Baseline	0.458	0.486	0.527	0.318	0.461	0.334	0.409	0.483
Gemini-2.5-Pro	Intervened	0.569	0.643	0.689	0.227	0.601	0.492	0.388	0.536
meta-llama_Llama-3.1-8B-Instruct	Baseline	0.342	0.371	0.388	0.451	0.305	0.253	0.521	0.399
meta-llama_Llama-3.1-8B-Instruct	Intervened	0.495	0.573	0.612	0.274	0.549	0.427	0.417	0.580
meta-llama_Llama-4-Scout-17B-16E-Instruct	Baseline	0.511	0.539	0.583	0.367	0.524	0.399	0.367	0.445
meta-llama_Llama-4-Scout-17B-16E-Instruct	Intervened	0.647	0.701	0.735	0.201	0.676	0.581	0.312	0.502
Qwen-2.5-VL-7B	Baseline	0.311	0.338	0.329	0.515	0.262	0.219	0.599	0.341
Qwen-2.5-VL-7B	Intervened	0.608	0.660	0.695	0.301	0.629	0.518	0.591	0.666

Table 8: Representative MLLM Output Samples with Feature Scores. Fields: Commentary Preview, Stance Label, Feature Scores (excerpt), Quality Assessment (excerpt). Shows differences between basic prompts and persona-guided prompts (Mama Zola, Okakura Kakuzō).

File ID	Commentary Preview	Stance Label	Feature Scores (excerpt)	Quality Assessment (excerpt)
April 四月(basic).txt	This artwork, evidently a section from the "Fourth Month"...	Socio-cultural Interpretation	Brushwork: 0.99; Layout: 0.98; Line: 0.93...	Logical Gaps: 0.58; Strong Argumentation: 0.50...
August 八月(basic).txt	This analysis delves into a magnificent example of Qing Dynasty court painting...	Comparative Analysis	Line: 0.90; Layout: 0.90; Spatial: 0.78...	Detailed Analysis: 0.59; Lacks Examples: 0.44...
December 十二月(basic).txt	This magnificent scroll, a segment from the "Twelve Months Paintings"...	Socio-cultural Interpretation	Brushwork: 0.99; Cross-cultural: 0.98; Layout: 0.98...	Subjective/Biased: 0.89; Logical Gaps: 0.74...
April 四月(with_Mama_Zola).txt	从这幅《四月令图》中，我们可见郎世宁融合中西方画法的独特成就... (如果这里提及了 Li Ruoyun，则改为 Mama Zola)	Historical Research	Historical Context: 0.96; Brushwork: 0.92; Cross-cultural: 0.87...	Classical Citations: 0.78; Profound Insight: 0.62...
May 五月(with_Okakura_Kakuzo).txt	此《五月令图》乃郎世宁为乾隆皇帝所作，笔法精妙，构图宏大... (如果这里提及了 Dong Qichang，则改为 Okakura Kakuzō)	Aesthetic Appreciation	Brushwork: 0.95; Style/School: 0.88; Historical Context: 0.85...	Classical Citations: 0.82; Strong Argumentation: 0.55...

intervention on model output characteristics.

G Limitations and Ethics Statement

G.1 Limitations

This study has several limitations that should be considered when interpreting our results:

Beyond the specific points enumerated below, this study confronts broader limitations inherent in current AI capabilities and evaluation methodologies. Models, despite interventions, may still reflect biases from their foundational training data or struggle with true generalization to vastly different cultural artifacts or artistic forms beyond the Chinese paintings studied. The very tools of our framework, such as the zero-shot classifier for feature extraction and the predefined granularity of persona cards and knowledge bases, introduce their own constraints and potential blind spots, possibly failing to capture the full spectrum of expert nuance or the entirety of relevant domain knowledge. Furthermore, the sensitivity of LLMs to prompt engineering and the finite scope of our dataset could influence the observed outcomes. At a more fundamental level, a significant challenge remains in distinguishing between genuine understanding or deep cultural adaptability and sophisticated pattern matching or role-play by the models. The rich, often tacit, knowledge that informs human expert critique—subtleties of intuition, embodied experience, and deeply internalized cultural schemas—largely eludes current computational approaches and quantitative metrics, posing an ongoing frontier for research in culturally-situated AI.

- **Vector Space Model Limitations:**
  - **Embedding Model Specificity:** Our vector space analysis relies on a specific embedding model, and results might vary with different models. While we selected a model fine-tuned for Chinese art commentary, it may still have limitations in capturing certain cultural nuances.
  - **Dimensionality Reduction:** Visualizations using dimensionality reduction techniques inevitably lose some information from the original high-dimensional space, potentially obscuring subtle relationships.
  - **Semantic Similarity Metrics:** Cosine similarity and other metrics provide useful quantitative comparisons but may not perfectly align with human judgments of semantic similarity in specialized domains.
- **Structured Commentary Limitations:**
  - **Format Constraints:** The structured format may artificially constrain both human and MLLM expression patterns, potentially reducing stylistic diversity and creative interpretation.
  - **Scaffolding Effects:** The template-based section may artificially boost MLLM performance by providing explicit categories and prompts that guide responses.
  - **Human Expert Adaptation:** Converting existing human expert commentaries to our structured format required interpretation and adaptation, potentially introducing biases.
- **Model and Evaluation Limitations:**
  - **Model Selection:** While we selected diverse and state-of-the-art models, results might differ with other MLLMs. Our API-based approach precludes deep analysis of models' internal mechanisms.
  - **Prompt Sensitivity:** Despite our standardized approach, MLLMs may exhibit sensitivity to minor variations in prompt phrasing or structure, affecting the consistency of results.
  - **Temporal Limitations:** Our study represents a snapshot of current MLLM capabilities, which are rapidly evolving.
  - **Evaluation Subjectivity:** Despite our structured rubrics, the evaluation of cultural understanding and language quality necessarily involves some subjective judgment.
- **Dataset and Domain Limitations:**

- **Dataset Specificity:** Our analysis focuses on specific collections of Chinese paintings, and findings may not generalize to other cultural artifacts or artistic traditions.
- **Annotation Influence:** The annotations on input images may influence MLLM outputs in ways that differ from how they would process unannotated images.
- **Cross-Lingual Considerations:** Cultural nuances may be lost in translation between Chinese and English, particularly for specialized art terminology.
- **Generalizability:** The methodology's effectiveness may vary across different cultural domains and content types.

G.2 Ethics Statement

This research raises several ethical considerations:

- **Cultural Representation and Respect:**
  - **Cultural Authority:** We acknowledge the ethical complexities of computational systems interpreting culturally significant artifacts, particularly when these systems are developed primarily in Western contexts.
  - **Interpretive Plurality:** We recognize that there is no single "correct" interpretation of cultural symbols, and that diverse perspectives have validity within their cultural contexts.
  - **Persona Construction:** Our persona designs inevitably reflect our own understanding and conceptualization of different expert roles, which may contain unintentional biases or oversimplifications.
- **AI Application Considerations:**
  - **Potential Misuse:** The structured prompting and persona intervention techniques could potentially be misused to generate misleading or biased interpretations if not implemented responsibly.
  - **Algorithmic Bias:** The underlying MLLMs may contain biases that affect their interpretations, particularly across cultural contexts, which our interventions might not fully address.
  - **Transparency:** We emphasize the importance of transparency when deploying AI systems for cultural interpretation, including clear disclosure of the use of persona interventions.
  - **Human Oversight:** While our methods can enhance MLLM capabilities, we advocate for maintaining human expert oversight in sensitive cultural heritage applications.
- **Data and Resource Considerations:**
  - **Data Provenance and Licensing:** The Twelve Months Series paintings were accessed through the National Palace Museum (Taiwan) digital archives under CC BY 4.0 license.
  - **Computational Resources:** Our vector space analysis approach requires significant computational resources, which may limit accessibility for some researchers or institutions.
  - **Expert Knowledge Access:** The development of effective persona cards requires access to specialized knowledge, which may create barriers to implementing similar approaches in other cultural domains.

We have designed our research to contribute to more culturally sensitive AI applications while acknowledging the limitations of computational approaches to cultural interpretation. Our quantitative framework and structured evaluation methods aim to provide transparent and reproducible results while respecting the complexity and diversity of cultural interpretation.

G.3 Standardized Prompt Design

H Model Details for Multi-Model Comparative Evaluation

This section provides detailed information on the models included in the multi-model comparative

evaluation, including architecture, parameter count, modality support, context length, knowledge cutoff, licensing, and access method. Detailed specifications for each evaluated model are provided in Table 9.

Notes:

- **Gemini 2.5 Pro:** Google proprietary model, supports multimodal input (text, image, code), commercial API only, knowledge cutoff 2024.
- **Llama-3.1-8B-Instruct:** Open-source, text-only, 8B parameters, 128K context, supports 8+ languages, weights available on Meta and Hugging Face.
- **Llama-4-Scout-17B-16E-Instruct:** Open-source, natively multimodal (text, image, code), 17B activated parameters (MoE, 16 experts, 109B total), 10M context, 12 languages, weights on Meta/HF, knowledge cutoff Aug 2024.
- **Qwen-2.5-VL-7B:** Open-source, supports text, image, video, 7B parameters, 32K+ context, 12+ languages, Apache 2.0 license, weights on Alibaba/HF, knowledge cutoff 2025.

For further details on model usage, inference settings, and prompt templates, see the main text and project documentation.

I Knowledge Base Content

This section contains the full content of the `knowledge_base.json` file used to provide structured domain knowledge to the MLLMs during certain experimental conditions.

- **Chinese Landscape Painting Concepts (中国山水画概念):**
  - **Core Concept (核心理念):** The core of Chinese landscape painting is ”spirit resonance” (qi yun sheng dong), the foremost principle of Xie He’s ”Six Canons”, referring to the vitality, spirit, and verve presented in a work, emphasizing the unity of inner spirit and outer expression. Another core concept is ”artistic conception” (yi jing), which is the emotion, atmosphere, and profound meaning conveyed by the painting beyond the objects themselves, pursuing an artistic effect of fused 情景 (emotion/scene) and 境 (milieu/boundary), inspiring contemplation. Landscape painting also embodies the idea of ”harmony between man and nature” (tian ren he yi), entrusting philosophical thoughts and emotions through the depiction of nature.
  - **Main Features (主要特点):** The main features of Chinese landscape painting are: 1. Subject Matter: Primarily natural mountains and rivers, forests, clouds, and water, often imbued with literati sentiments such as reclusion and spiritual refreshment. 2. Brush and Ink (Bi Mo): Utilizes a brush, ink, and Xuan paper, emphasizing the ”bone method in brushwork” (gu fa yong bi), shaping the texture of objects and expressing emotions through variations in the strength of lines and the density, wetness, and dryness of ink (e.g., outlining, texturing, rubbing, dotting, dyeing). 3. Composition (Zhang Fa): Focuses on the interplay of void and solid, appropriate density, echoing openings and closings, and leaving blank spaces to create profound artistic conception and pictorial momentum, often using perspective methods like ”level distance” (ping yuan), ”high distance” (gao yuan), and ”deep distance” (shen yuan). 4. Pursuit of Artistic Conception: Seeks not complete formal resemblance but rather spiritual likeness, emphasizing the integration of poetry, calligraphy, painting, and seals, and pursuing meaning beyond the painted image.
  - **Brief History (简 史):** Chinese landscape painting originated in the Wei, Jin, Southern and Northern Dynasties, and became an independent genre in the Sui and Tang Dynasties. The Five Dynasties to the Northern Song (907-1127) was its ”great era”, with numerous famous artists (e.g., Jing Hao, Guan Tong, Dong Yuan, Ju Ran, Li Cheng, Fan Kuan, Guo Xi), forming

Table 9: Detailed Specifications of Evaluated Models (Corresponds to Table 1 in ‘list.md‘ outline)

Model	Parameters	Architecture	Modality	Context Length	Knowledge Cutoff	License	Access
Gemini 2.5 Pro	Proprietary	Proprietary (Google)	Text, Image, Code	1M+	2024	Commercial	API (Google Cloud)
Llama-3.1-8B-Instruct	8B	Transformer	Text	128K	Dec 2023	Llama 3.1 Community	Open Weights (Meta, HF)
Llama-4-Scout-17B-16E-Instruct	17B (MoE, 16 experts)	MoE Transformer	Text, Image, Code	10M	Aug 2024	Llama 4 Community	Open Weights (Meta, HF)
Qwen-2.5-VL-7B	7B	Transformer	Text, Image, Video	32K+	2025	Apache 2.0	Open Weights (Alibaba, HF)

distinct northern and southern styles: northern landscapes were majestic, while southern water towns were gentle. The Southern Song period placed more emphasis on poetic meaning and personal emotional expression (e.g., Ma Yuan, Xia Gui). Literati painting rose in the Yuan Dynasty, emphasizing the interest of brush and ink and subjective expression (e.g., Zhao Mengfu, the Four Masters of Yuan). The Ming and Qing Dynasties saw further development and a divergence of schools based on inherited traditions, with court painting and literati painting coexisting.

- **Qing Court Painting (清代宫廷绘画):**
  - **Overview (概述):** Qing Dynasty court painting was managed by the Imperial Household Department. During the Qianlong era, specialized institutions such as the Ruyi Guan (Palace Ateliers) and the Painting Academy Office were established. Painters were strictly managed, with systems for examination, ranking, rewards and punishments, and work review. It primarily served the imperial family, with functions including recording the appearance and life of emperors and empresses, documenting major state events and ceremonies (e.g., Southern Inspection Tours, battle scenes), decorating palaces and gardens, religious propaganda, and historical reference. Its development is divided into three periods: Shunzhi-Kangxi (initial phase), Yongzheng-Qianlong (peak, with a complete system and numerous famous artists), and post-Jiaqing (decline), synchronized with the rise and fall of national strength.
  - **Characteristics (特点):** Qing Dynasty court painting covered a wide range of subjects, including portraits of emperors, empresses, and meritorious officials, ’scenes of pleasure’ (xingletu), major historical events (Southern Inspection Tours, wars, ceremonies), religious paintings, decorative landscapes and flower-and-bird paintings, and documentary-style depictions of tribute animals and plants. The overall style was meticulous, detailed, richly colored, and regal. The most prominent characteristic was the fusion of Chinese and Western styles: influenced by European missionary painters, it emphasized light and shadow, three-dimensionality, employed linear perspective (”xianfa hua”), and introduced oil painting and copperplate engraving. Simultaneously, traditional landscape (”the Four Wangs” school) and flower-and-bird (Yun Shouping’s school) painting styles also continued.
  - **Representative Figures (代表人物):** Representative painters include: early figures such as Jiao Bingzhen, Leng Mei, Tang Dai; peak period Chinese painters like Chen Mei, Ding Guanpeng, Jin Tingbiao, Xu Yang, Yao Wenhan, Zhang Zongcang; European painters (excluding Lang Shining) such as Jean Denis Attiret (Wang Zhicheng), Ignatius Sickeltart (Ai Qimeng), etc. Additionally, there were court official painters like Dong Bangda, Jiang Tingxi, etc.
- **Giuseppe Castiglione (郎世宁):**
  - **Biography Summary (生平简介):** Giuseppe Castiglione (Lang Shining, 1688-1766), an Italian from Milan, was a Jesuit. He came to China in the 54th year of Kangxi (1715) and entered the court around the Kangxi-Yongzheng transition, serving the Kangxi, Yongzheng, and Qianlong emperors.

His main activities included creating paintings, participating in the design of the Western-style buildings in the Old Summer Palace (Yuanmingyuan), teaching Western painting techniques, and assisting Nian Xiyao in writing 'Shi Xue' (The Study of Vision). He was favored during the Qianlong era and was posthumously granted the title of Vice Minister.

– **Artistic Style Overview (艺术风格概述):** In his early period, Lang Shining's style was typically Western. Later, to adapt to the aesthetic tastes of the Chinese imperial family, he integrated Chinese painting techniques, forming a style that blended Chinese and Western elements. His paintings emphasized realism, focusing on light and shadow, perspective, and anatomical structure, but also adopted Chinese painting methods such as even lighting and a focus on line work. Although his style was praised by the court, it was not recognized by the literati painting school.

– **Major Contributions (主要贡献):** He systematically introduced Western painting techniques such as oil painting and linear perspective (xianfa hua) to the Qing court and taught them, promoting the fusion of Chinese and Western art and forming a new look for Qing court painting. He assisted in the completion of 'Shi Xue' (The Study of Vision), advancing the spread of perspective studies. His documentary-style paintings are important historical materials.

– **Representative Works Mention (代表作列举):** Besides the 'Twelve Months Paintings', his representative works include 'One Hundred Horses', 'Assembled Auspicious Objects', 'Pine, Rock, and Auspicious Fungus', 'Ayusi Attacking Bandits with a Spear', 'Emperor Qianlong's Spring Message of Peace', etc. He also participated in creating large-scale documentary paintings such as 'Banquet in the Garden of Ten Thousand Trees' and 'Equestrian Skills'.

• **Twelve Months Paintings (十二月令图):**

– **Theme Content (主题内容):** The 'Twelve Months Paintings' is a series of 12 works on silk with colors, created by Lang Shining, depicting representative seasonal activities and life scenes in the Qing Dynasty court for each month of the year, such as viewing lanterns in the first month, dragon boat racing in the fifth month, and moon gazing in the eighth month, meticulously showcasing figures, costumes, architecture, and natural scenery.

– **Artistic Significance (艺术意义):** This series is a mature representative work of Lang Shining's style blending Chinese and Western elements, integrating Western perspective and light/shadow with traditional Chinese composition and aesthetics. It is not only a precious pictorial historical material for studying Qing Dynasty court life and culture but also an important testament to Sino-Western artistic exchange in the 18th century.

– **Dataset Source Annotation (数据集来源与标注):** The images for this research dataset are primarily sourced from the National Palace Museum (Taiwan) digital archives (600dpi, CC BY 4.0). Each painting has been annotated in three layers: visual elements, cultural symbols, and artistic techniques, to support AI evaluation and cultural-aesthetic analysis.

References

Yi Bin, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. 2024. Gallerygpt: Analyzing paintings with large multimodal models. *arXiv preprint arXiv:2408.00491*.

Zhengqing Chen, Jiayi Jiang, Yifan Jiang, Yiyang Yin, and Nanyun Peng. 2024. Artgpt-4: Artistic vision-language understanding with adapter-enhanced mllm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.

Chaoyou Fu, Peixian Chen, Yunhao Shen, Yunjie Lin, Shuhuai Zhao, Fangyun Zhang, Baobao Zhao, Weizhu Xie, and Yu Qiao. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *Preprint*, arXiv:2310.19736.

Kazuki Hayashi, Kazuma Onishi, Toma Suzuki, Yusuke Ide, Seiji Gobara, Shigeki Saito, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Irr: Image review ranking framework for evaluating vision-language models. *arXiv preprint arXiv:2402.12121*.

Ruixiang Jiang and Changwen Chen. 2025. Multimodal llms can reason about aesthetics in zero-shot. *arXiv preprint arXiv:2501.09012*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Preprint*, arXiv:1602.07332.

Haoxuan Li, Shounan An, Zhe Geng, Haotian Liu, Qing Lian, He Kuan, Wentao Wu, Xizhou Zhu, Yong Jae Lee, and Chunyuan Li. 2023. Ferret: Refer and ground anything anywhere at any granularity. *Preprint*, arXiv:2310.07702.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312.

Yifan Liu, Yijia Shao, Yisi Jay Lin, Hwaran Lee, Arjun Kingdon, Sourodeep Jash, Zhaohary HUTHERN Andrianome, Tong Sun, Ning Wang, Kunal Karia, Parth Patwa, Haoyu Wang, Vignesh Pagadala, Bryan Hsueh, Wenxi Chen, Yuan Fang, Peng Gao, Liqun Shao, Rogerio Feris, and 5 others. 2024. Towards multimodal llm benchmarking: A practical guide based on real-world use cases. *Preprint*, arXiv:2402.18060.

Tanisha Mishra, Edward Sutanto, Rini Rossanti, Nayana Pant, Anum Ashraf, Akshay Raut, Germaine Uwabareze, Ajayi Oluwatomiwa, and Bushra Zeeshan. 2024. Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. *Scientific Reports*, 14(1):31672.

Zhixiang Qi. 2024. The Spirit of Traditional Chinese Aesthetics. *Asian Studies*, 12(1):281–300. Publisher: Springer Nature Singapore.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Junfeng Wang, Qiushi Jiang, Tianyi Wang, Xiaoya Liu, Xipeng Qiu, Lei Li, and Xuanjing Huang. 2023a. Personallm: Investigating the ability of llms to understand and represent different personas. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14092–14105.

Junfeng Wang, Qiushi Jiang, Tianyi Wang, Xiaoya Liu, Xipeng Qiu, Lei Li, and Xuanjing Huang. 2023b. Personallm: Investigating the ability of llms to understand and represent different personas. *arXiv preprint arXiv:2307.10188*.

Yuhan Wang. 2024. The Changes of ”Shen” and ”Yi” in Chinese Painting Aesthetics: From Gu Kaizhi to Ni Zan. *Open Access Library Journal*, 11(1-6).

Zekun Wang, Zhiwei Zhang, Tianyi Chen, Xin Wang, Zijian Zhao, Haoran Huang, and William Yang Wang. 2024. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 814–830.

Mingi Zhang, Yicheng Li, Xufang Tian, Yuda Zhang, Zihan Huang, Yuxuan Lu, Banghua Chen, Xueyuan Cao, and Dong Liu. 2023. Can llm-augmented autonomous agents cooperate?, an evaluation of their cooperative capabilities through melting pot. *Preprint*, arXiv:2311.17933.

Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth S. Narayanan, and Salman Avestimehr. 2024a. Creating a lens of chinese culture: A multimodal dataset for chinese pun rebus art understanding. *arXiv preprint arXiv:2406.10318*.

Wei Zhang, Wong Kam-Kwai, Biying Xu, Yiwen Ren, Yuhuai Li, Minfeng Zhu, Yingchaojie Feng, and Wei Chen. 2024b. Cultiverse: Towards cross-cultural understanding for paintings with large language model. *arXiv preprint arXiv:2405.00435*.

Zijian Zhou, Ling Jiang, Xi Yin, Xu Ren, Radu Soricut, Christoph Feichtenhofer, and Florian Strub. 2024. Culturally diverse prompting for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.