AllCuration: Toward Integrated Art Curation with Multimodal Large Language Models

Anonymous ACL submission

Abstract

Existing multimodal approaches to art analysis often focus on isolated aspects, such as visual features, emotions, or cultural context, leading to fragmented interpretations. To address this limitation, this study proposes an integrated framework that jointly considers the formal, emotional, and cultural dimensions of artworks. To this end, we construct the **ArtCura** dataset by extending EmoArt with culturally grounded analyses generated by GPT-4o and refined using SelfCheckGPT to improve reliability. Based on this dataset, we further develop **AllCuration**, a multimodal large language model fine-tuned on a ShareGPT4V backbone to generate comprehensive art critiques. Experimental results show that the proposed model outperforms existing multimodal baselines, and human evaluations demonstrate that it produces more balanced and contextually grounded analyses. These findings suggest that integrating formal elements, emotional responses, and cultural context is essential for generating effective art critiques. The code is available at https://anonymous.4open.science/r/AllCuration-D6C4/.

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Liu et al., 2023b; Li et al., 2023) have significantly improved the ability of AI systems to understand and describe visual content. These models have demonstrated strong performance in various vision–language tasks, including image captioning (Liu et al., 2023a; Ramakrishnan et al., 2025), visual question answering (Antol et al., 2015), and multimodal reasoning (Huang et al., 2023; Zhou et al., 2025). Building on these developments, recent studies have begun to explore the application of multimodal models to art analysis and critique generation, aiming to support the interpretation and understanding of artworks (Hayashi et al., 2024).

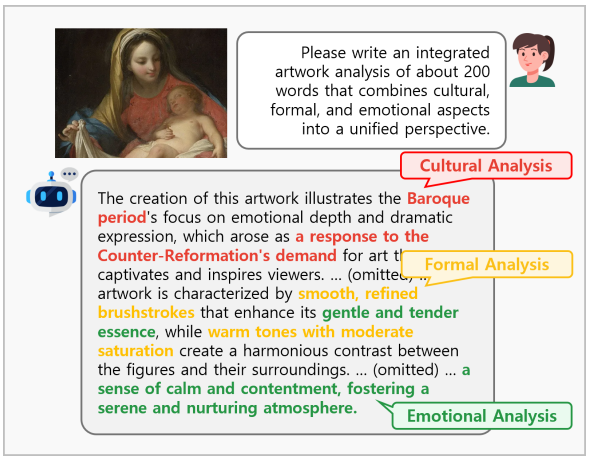

Figure 1: Illustration of AllCuration's integrated art analysis, where cultural (red), formal (yellow), and emotional (green) dimensions are coherently combined into a single, unified analysis.

GalleryGPT (Bin et al., 2024) focuses on generating artwork descriptions based on formal visual elements such as color, composition, and lighting. In addition, recent VLM-based artwork interpretation studies have primarily examined specific interpretive aspects, including emotional and symbolic interpretations of artworks (Padó and Thomas, 2025). Several datasets have been proposed to support art understanding. ArtEmis (Achlioptas et al., 2021) collects emotional reactions and natural language descriptions of artworks, enabling an analysis of the relationship between visual art and emotional expression. EmoArt (Zhang et al., 2025) introduced a multidimensional art dataset that incorporates emotion categories and visual attributes, such as color and composition. Furthermore, AQUA (Garcia et al., 2020) provides visual question answering data combined with art historical knowledge to address the contextual understanding of artworks.

However, these studies tend to treat the formal,

emotional, and cultural dimensions independently. In the field of art criticism, it has been argued that meaningful interpretation requires the integration of multiple interpretive dimensions rather than isolated approaches to interpretation. In particular, Noël Carroll's theory of art criticism (Carroll, 2009) conceptualizes art criticism as an integrated process in which description, analysis, contextualization, interpretation, and evaluation are closely interlinked. Within this framework, description and analysis address the perceptual and formal structure of an artwork, contextualization incorporates historical and cultural backgrounds, and interpretation and evaluation consider the meanings and emotional effects conveyed by the work. This perspective suggests that meaningful art interpretation requires not reliance on a single dimension but an integrated understanding that connects formal perception, emotional response, and cultural context.

Based on this perspective, we propose an integrated framework for multimodal art curation and critique generation. First, we construct ArtCura, an instruction dataset that extends the EmoArt dataset by incorporating cultural analyses generated by GPT-4o, thereby enabling the integration of the cultural, formal, and emotional dimensions of artworks. To improve the reliability of the generated cultural analyses, we further apply SelfCheckGPT (Manakul et al., 2023) to mitigate hallucinations in the text generation process. Finally, we propose AllCuration, a ShareGPT4V-based model fine-tuned on the ArtCura dataset, which generates comprehensive and in-depth art critiques, as illustrated in Figure 1, by integratively incorporating formal elements, emotional responses, and cultural context. The main contributions are presented as follows:

- We construct the **ArtCura** dataset by extending EmoArt with GPT-4o-generated cultural analyses, making it one of the first multimodal art datasets to integrate cultural, formal, and emotional dimensions.
- To improve the reliability of the generated cultural analyses, we apply SelfCheckGPT to mitigate hallucinations during text generation.
- We propose **AllCuration**, a ShareGPT4V-based model fine-tuned on ArtCura for generating comprehensive art critiques that integrate formal, emotional, and cultural aspects.

2 Related Work

2.1 MLLMs for Art Analysis

Recent advances in Multimodal Large Language Models (MLLMs) (Antol et al., 2015; Liu et al., 2023a; Ramakrishnan et al., 2025) have significantly increased interest in artwork analysis and critique generation. GalleryGPT (Bin et al., 2024) utilizes instruction-based image–text pairs to generate formal analyses of paintings, demonstrating that visual elements such as color, composition, and lighting can enhance artwork understanding. In addition, recent studies on MLLM-based aesthetic reasoning have advanced toward generating aesthetic evaluations and interpretations that resemble human artistic judgment through evidence-grounded reasoning processes (Jiang and Chen, 2025). Furthermore, recent VLM-based artwork interpretation studies have focused on specific interpretive dimensions, such as emotional and symbolic interpretations of artworks (Padó and Thomas, 2025). These studies analyze the capability of VLMs to interpret emotional expressions and symbolic elements in artworks, highlighting the importance of emotional and symbolic aspects in art understanding. However, because these approaches focus on individual interpretive aspects, a more comprehensive understanding of artworks requires an integrated approach that jointly considers formal, emotional, and cultural dimensions.

2.2 Datasets for Multimodal Art Understanding

Existing datasets for multimodal art understanding primarily focus on specific interpretive dimensions, such as emotion, visual elements, and contextual knowledge. ArtEmis (Achlioptas et al., 2021) is a large-scale dataset that collects emotional responses to artworks and provides emotion labels and natural language descriptions to emphasize the subjectivity and diversity of emotional interpretation. EmoArt (Zhang et al., 2025) extended this direction by providing multidimensional information, including visual attributes, content descriptions, and emotional annotations, enabling the analysis of the relationships between formal visual elements and emotional responses to them. Meanwhile, AQUA (Garcia et al., 2020) introduced a visual question answering dataset that distinguishes between visual and knowledge-based questions, highlighting the importance of art-historical knowledge and contextual understanding.

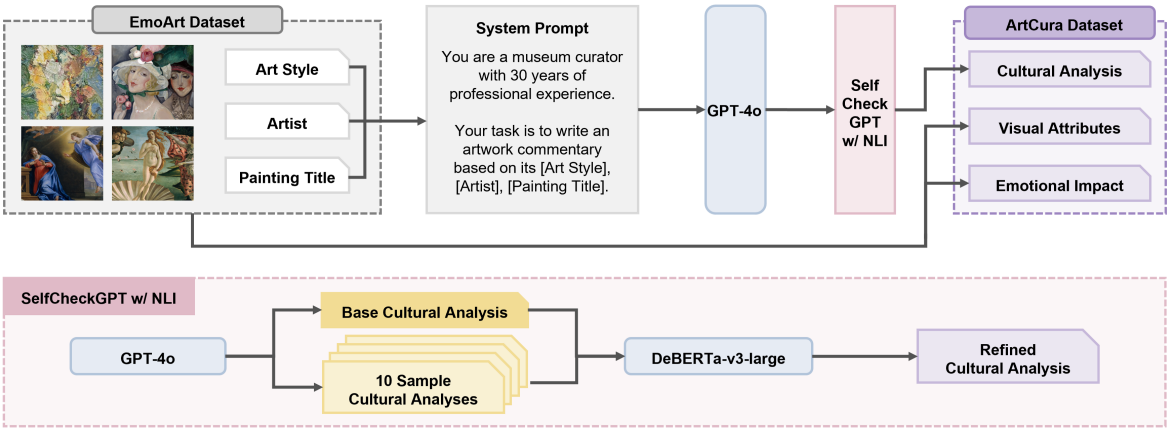

Figure 2: ArtCura dataset construction pipeline. GPT-4o generates cultural analyses from artwork metadata, such as art movements, artists, and painting titles. SelfCheckGPT with NLI-based verification improves the consistency and factual reliability of the generated analyses. Finally, the refined cultural analyses are integrated with visual attributes and emotional annotations to construct the ArtCura dataset for multidimensional artwork understanding.

However, these datasets mainly focus on emotional, visual, or knowledge-based aspects independently and do not sufficiently support interpretations that integratively reflect the formal, emotional, and cultural dimensions. Therefore, a more comprehensive understanding of artwork requires datasets that jointly consider multiple interpretive dimensions.

2.3 Mitigating Hallucination in LLM-Generated Content

Although large language models (LLMs) demonstrate strong text generation capabilities, hallucination, which produces factually incorrect information, remains a major limitation (Maynez et al., 2020; Ji et al., 2023; Hayashi et al., 2024). Moreover, automatic evaluation metrics such as BLEU and ROUGE fail to adequately assess the factual accuracy of the generated outputs (Kryściński et al., 2020; Gabriel et al., 2021).

To address this issue, SelfCheckGPT (Manakul et al., 2023) was proposed as a method for evaluating the factuality of LLM-generated texts without relying on external data or ground-truth annotations. This is based on the principle that information consistently appearing across multiple responses generated from the same prompt is more likely to be factual. In this study, we apply SelfCheckGPT during the dataset construction process to verify the cultural background information of artworks, thereby improving the factual reliability of cultural analyses without costly manual verification.

3 Proposed Method

3.1 Cultural Analysis Generation

To achieve a comprehensive understanding of artworks, it is necessary to consider not only formal characteristics but also emotional expression and cultural context. Visual elements such as color, composition, and brushwork convey external characteristics, but their meanings become clearer when interpreted in relation to the socio-historical context of an artwork. Therefore, in this study, we construct the ArtCura dataset by extending EmoArt (Zhang et al., 2025) with cultural analysis information, as illustrated in Figure 2.

The EmoArt dataset collects artwork images from publicly available sources such as WikiArt and the Metropolitan Museum of Art and provides annotations across three dimensions: (1) *Content Description*, (2) *Visual Attributes with corresponding Emotional Impact*, and (3) *Emotional Categories*. While this structure enables the analysis of formal and emotional aspects, it does not include the cultural information necessary to explain how artworks are shaped by specific social and historical contexts.

To address this gap, we use the GPT-4o large language model to generate cultural analyses of 5,132 artworks. Because this cultural analysis generation relies on external knowledge sources, such as artists' biographies and socio-historical contexts, we exclude 324 artworks with incomplete or ambiguous metadata. In addition, we design a rule-based prompt to generate cultural analyses, where

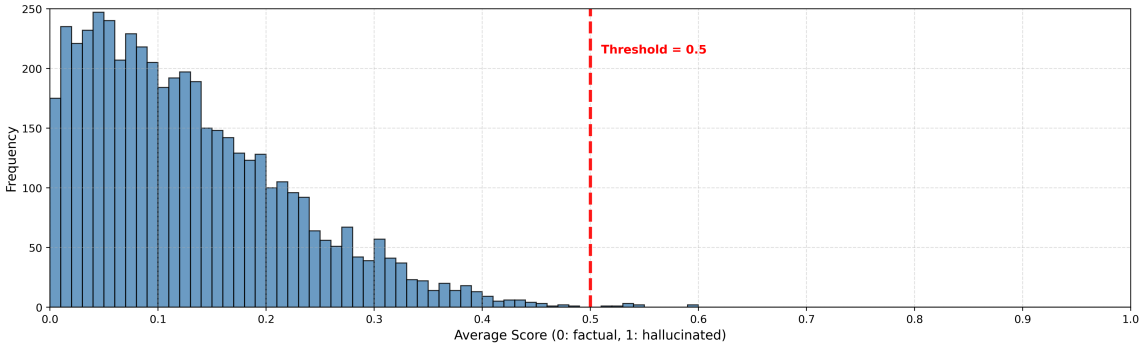

Figure 3: Distribution of SelfCheck NLI average scores

the model is instructed to act as an experienced museum curator and base its analysis on art-historical knowledge. Through this process, GPT-4o generates cultural analyses describing how each artwork reflects its cultural and historical contexts.

3.2 Dataset Refinement and Construction

Large language models, such as GPT-4o, can produce hallucinations during cultural analysis generation. To mitigate this issue, we applied the SelfCheckGPT method (Manakul et al., 2023), which evaluates the reliability of the generated content at the sentence level using Natural Language Inference (NLI).

For each cultural analysis (CA), we generated 10 supplementary samples ($S_1, S_2, \ldots, S_{10}$) using the same prompt. The target analysis was generated at a temperature of 0 to maintain consistency, while the supplementary samples were generated at a temperature of 1 to introduce controlled variation. Each sentence (CA_i) in the original analysis was evaluated by treating S_k as the premise and CA_i as the hypothesis in the NLI model.

The NLI model classifies sentence pairs into entailment, neutral, and contradiction. We use the contradiction probability as the SelfCheck score (Manakul et al., 2023). Formally, the SelfCheckGPT-NLI score for each sentence CA_i is defined as:

$$S_{NLI}(i) = \frac{1}{K} \sum_{k=1}^{K} P(\text{contradict} \mid CA_i, S_k) \quad (1)$$

A higher score indicates lower semantic consistency, implying that the sentence is more likely to contain hallucinations. Figure 3 presents the distribution of the average SelfCheckGPT-NLI scores computed over the dataset. The distribution is concentrated in the low-score region and exhibits a right-skewed pattern, with a mean of 0.1291 and a standard deviation of 0.0945. Based on this distribution, we used a threshold of $\tau = 0.5$ and removed the sentences that exceeded this value. The details of the threshold selection are provided in Section 5.3.

After refinement, the final ArtCura dataset comprises 4,799 images with culturally and historically grounded analyses. We split the refined dataset into training and test sets with a ratio of 8:2, resulting in 3,839 training samples and 960 test samples. To compensate for the reduced dataset size, we applied data augmentation to the training set by generating three instruction-paraphrased variants per sample, resulting in 11,517 training instances. The test set was not augmented to ensure fair evaluation.

3.3 AllCuration: Supervised Fine-Tuning

To verify whether the ArtCura dataset can enhance the capability of MLLMs in generating comprehensive art critiques, we developed AllCuration. We perform supervised fine-tuning (SFT) with LoRA (Hu et al.) on the ShareGPT4V (Chen et al., 2024) backbone using ArtCura for parameter-efficient adaptation, which enables the model to specialize in the art critique domain without full parameter updates, reducing computational cost while maintaining performance. The SFT process utilizes instruction–answer pairs to align each painting with its corresponding formal, emotional, and cultural analyses.

Unlike existing vision–language models that primarily emphasize formal visual features, AllCuration aims to generate integrated artwork interpretations by jointly considering the formal, emotional, and cultural dimensions. Through fine-tuning on

Table 1: Survey items used in the human evaluation (5-point Likert-scale)

Category	ID	Survey Item
Cultural	Q1	The description appropriately reflects the historical background and artistic movement of the artwork (e.g., Renaissance, 1990s).
	Q2	The description clearly presents societal issues or historical events that influenced the work (e.g., World Wars, digital revolution).
	Q3	The description effectively expresses the cultural or temporal atmosphere of the period.
Formal	Q4	The description explains the formal structure of the artwork, including its composition, balance, and spatial layout.
	Q5	The description reflects the mood and depth through the use of color, contrast, and lighting.
	Q6	The description captures visual features such as brushstrokes and line quality.
Emotional	Q7	The description expresses the emotions conveyed by the artwork (e.g., sadness, anger or joy).
	Q8	The description captures the emotional tone or atmosphere of the artwork (e.g., calm, mysterious or captivating).
	Q9	The description uses emotional language effectively to convey the affective tone (e.g., serene, anxious, and warm).
Integrated	Q10	The description provides a well-balanced integration of cultural, formal, and emotional perspectives.

the ArtCura dataset, the model learns to associate paintings with multidimensional analyses, enabling context-aware and comprehensive art critiques.

4 Experiments

4.1 Experimental Setup

We compared AllCuration with several open-source multimodal baselines, including LLaVA-1.5 (Liu et al., 2023b), ShareGPT4V (Chen et al., 2024), Qwen-VL-Chat (Bai et al., 2023), Qwen2.5-VL (Bai et al., 2025), and GalleryGPT (Bin et al., 2024). AllCuration was fine-tuned on an RTX 6000 GPU with a learning rate of 2e-5 and a batch size of 16. For the LoRA module, the LoRA rank and alpha values were set to 128 and 256, respectively.

A total of 960 paintings from the SelfCheckGPT-validated dataset were used as the test set to assess the model accuracy. All models were evaluated under the same instruction prompt, and to ensure consistency, all outputs were generated at a temperature of 0.2, thereby minimizing randomness during evaluation.

4.2 Quantitative Evaluation Metrics

We employed standard image captioning metrics, including BLEU (Papineni et al., 2002), GLEU (Mutton et al., 2007), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and BERTScore (Zhang et al.). Artwork analysis can be formulated as an image-grounded description generation task that incorporates visual, cultural, and emotional elements. Accordingly, image captioning metrics are suitable for evaluating the quality of the generated descriptions and their semantic consistency with the reference texts. In particular, BERTScore is effective in capturing cultural and emotional meanings that may be expressed in diverse ways (Zhang et al.).

4.3 Qualitative Evaluation Setup

To evaluate the ability of multimodal large language models (MLLMs) to generate insightful and interpretable art criticism, we conducted a human evaluation comparing the proposed model with five baseline MLLMs, including LLaVA-1.5, ShareGPT4V, Qwen-VL-Chat, Qwen2.5-VL-7B, and GalleryGPT. Ten participants were randomly divided into two groups, each of which evaluated model-generated descriptions of five artworks. The participants compared the descriptions produced by different models for the same artwork, enabling direct comparisons under identical visual conditions. In total, ten paintings were evaluated, with each artwork assessed by multiple participants to reduce individual subjectivity and improve the reliability of the evaluation.

Each description was evaluated using a 5-point Likert scale across 10 items grouped into four categories based on the art criticism framework proposed by Carroll (2009), as summarized in Table 1. The Cultural category (Q1–Q3) assessed historical, artistic, and societal context; the Formal category (Q4–Q6) evaluated visual properties such as composition, color, and brushwork; the Emotional category (Q7–Q9) measured affective tone and emotional language; and the Integrated category (Q10) assessed whether cultural, formal, and

Table 2: Internal consistency reliability across evaluation categories (Cronbach's α)

Evaluation Category	Cronbach's α
Cultural (Q1–Q3)	0.839
Formal (Q4–Q6)	0.600
Emotional (Q7–Q9)	0.840

emotional perspectives were well balanced and coherently integrated.

To assess the internal consistency of the survey items, we computed Cronbach's α values for each category. The results in Table 2 show that all categories exhibit acceptable reliability, indicating that the items are coherently aligned with their intended constructs and can be aggregated into category-level scores. The Formal category showed relatively lower α values, which may be attributed to the subjective and diverse nature of formal artistic elements, such as composition, color usage, and brushstroke style. Because these items capture different perceptual aspects of artworks, they may exhibit relatively lower internal consistency than the other categories.

5 Results

5.1 Quantitative Results

Table 3 presents the quantitative comparison results based on the standard image captioning metrics. GalleryGPT shows relatively strong textual fluency, as reflected in its high GLEU score, whereas Qwen2.5-VL demonstrates a comparatively strong performance on METEOR. However, the BLEU and ROUGE scores across all models remain relatively low overall, as these metrics primarily rely on n-gram overlap, whereas artwork criticism is inherently open-ended and permits diverse interpretations and expressions, even for the same artwork. In contrast, AllCuration achieves the highest performance across all evaluation metrics, indicating more balanced overall results. These findings suggest that AllCuration effectively captures and integrates the formal, emotional, and cultural dimensions of artwork analysis. Furthermore, the combination of a SelfCheckGPT-validated dataset and supervised fine-tuning on the ShareGPT4V backbone appears to contribute to the generation of more comprehensive and integrated artwork analyses. In addition, AllCuration passed the hallucination check, indicating improved consistency in the generated descriptions.

Table 3: Performance comparison of artwork analysis generation measured by captioning metrics (%). B: BLEU, G: GLEU, M: METEOR, R: ROUGE, BS: BERTScore, HC: Hallucination Check

Model	B	G	M	R	BS	HC
LLaVA-1.5	2.63	11.68	27.24	20.70	86.38	Passed
ShareGPT4V	2.94	10.89	27.64	20.00	86.12	Passed
GalleryGPT	1.95	12.03	24.12	20.64	86.57	Passed
Qwen2.5-VL	3.02	9.33	31.97	18.86	86.29	Failed
Qwen-VL-Chat	2.01	9.59	26.62	19.32	85.40	Failed
AllCuration (Ours)	**10.84**	**20.09**	**39.09**	**30.87**	**90.28**	Passed

5.2 Qualitative Results

5.2.1 Hallucination Anlaysis

To further assess the factual reliability of the generated descriptions, we conducted a manual hallucination inspection of the top-10 samples per model ranked by BERTScore. The overall results are summarized in Table 3. Three independent evaluators examined each instruction–response pair for three types of factual errors: (1) temporal errors, (2) art-movement errors, and (3) artwork title errors. Each sample was evaluated independently by all three evaluators, followed by a final consolidation step conducted by a single reviewer to ensure inter-evaluator consistency and resolve any minor discrepancies.

The inspection results revealed distinct error patterns across the different models. GalleryGPT and LLaVA-1.5 mainly exhibited temporal errors, whereas ShareGPT4V showed more art movement-related errors. Hallucinations were identified in four out of ten cases for GalleryGPT, six for LLaVA-1.5, and five for ShareGPT4V. Qwen2.5-VL exhibited high hallucination rates across all error categories, often involving multiple factual errors within a single response. Qwen-VL-Chat exhibited the highest hallucination rate, primarily involving artist name and artwork title errors. Due to their high error frequency and severity, Qwen2.5-VL and Qwen-VL-Chat were excluded from subsequent human evaluation and statistical analysis.

5.2.2 Human Evaluation Results

Prior to examining the human evaluation results, we computed the intraclass correlation coefficient (ICC (2, k)) to quantitatively assess inter-rater reliability among the evaluators. ICC(2,k) measures reliability for the average scores of k raters under a two-way mixed-effects model and is appropriate for our design in which the mean scores of multiple raters are used as the final evaluation score (Koo

Academicism
Emile Munier
Playing with the Kittens
(1840 - 1895)

ShareGPT4V

This is a captivating oil painting that transports us back to the 19th century. The art style is reminiscent of the ***Romanticism movement***, (omitted) The painting captures a tender moment between a woman and two children, seated on a stone bench. (omitted) The warm tones of the colors evoke a sense of warmth and comfort, while the play of light and shadow adds depth and dimension to the painting. (omitted) In terms of emotional impact, the painting evokes feelings of nostalgia and warmth. The depiction of a simple, everyday moment is relatable and heartwarming, reminding us of the beauty and simplicity of life.

GalleryGPT

This idyllic painting depicts a serene family gathering in a sunlit garden. The composition is dominated by three figures: a young woman, a child, and an elderly man. The woman, clad in a white dress, stands in the foreground, (omitted) The soft, muted colors and diffused light create a harmonious and tranquil atmosphere, while the careful rendering of the figures and their surroundings conveys a sense of realism and intimacy.

LLaVa 1.5

The painting depicts a scene of a woman and two children sitting on a stone wall, enjoying a meal together. (omitted) The artwork is situated within the context of the 19th century, a time when family values and togetherness were highly valued. The social environment of the time emphasized the importance of family life and the role of women in caring for their children. (omitted) From a formal perspective, the painting is a realistic representation of the scene, with the woman and children depicted in a lifelike manner. The use of light and shadow adds depth and dimension to the image, (omitted) The painting captures a moment of love between the woman and her children, reflecting the social and emotional values of the 19th century. (omitted)

AllCuration (Ours)

The creation of this artwork illustrates the 19th-century fascination with the East, spurred by colonial expansion and a romanticized view of Eastern cultures. (omitted) This piece exemplifies the era's tendency to interpret Eastern themes through a Western perspective, often influenced by travel literature and the growing field of anthropology. The artwork is characterized by smooth, fine brushstrokes that enhance its gentle and harmonious quality. (omitted) by warm tones, with soft greens and browns that create a welcoming atmosphere. The composition is both balanced and dynamic, (omitted) while the natural lighting and soft shadows contribute to a warm, inviting ambiance. The delicate, smooth lines further emphasize the peaceful (omitted) warm color palette evoke a sense of calm and contentment, while the dynamic composition and soft lighting create an inviting and peaceful environment.

Figure 4: Qualitative comparison of artwork analyses generated by multimodal models. Pink highlights indicate cultural interpretations, blue highlights indicate visual elements, green highlights indicate emotional descriptions, and red highlight indicates art-movement hallucination discussed in Section 5.2.1. Compared to baseline models, our method produces more grounded and coherent artwork interpretations with fewer hallucinations.

and Li, 2016).

The ten participants were divided into two groups of five (Group A and Group B), each evaluating a different set of artworks. Group A yielded ICC(2,k) = 0.641, corresponding to a moderate level of agreement, indicating that raters generally agreed in their assessments, while some inter-individual variation remained. Group B yielded ICC(2,k) = 0.831, reflecting good agreement and consistent judgment criteria. The average across both groups was 0.736, indicating an acceptable level of inter-rater reliability.

The difference in ICC values between the two groups is a commonly observed phenomenon in human evaluation studies that employ small rater panels. In particular, the critical evaluation of artworks is inherently susceptible to individual variations in cultural background, art education, and personal aesthetic experience, which tend to yield lower inter-rater agreement than more objective measurement tasks. In this context, the obtained ICC(2,k) values are considered acceptable, given the subjective nature of art criticism, and are further supported by the adequate internal consistency of the survey items within each category.

The results of the human evaluation are summarized in Table 4. AllCuration achieved the highest mean scores across all four evaluation categories, outperforming all baseline models in the Cultural, Formal, Emotional, and Integrated dimensions. These results demonstrate that AllCuration consistently generates artwork analyses that are not only structurally and historically grounded but also affectively expressive and well integrated across multiple critical perspectives. This tendency can also be observed in the qualitative comparison results shown in Figure 4.

Table 4: Descriptive analysis of human evaluation scores

Model	Cultural	Formal	Emotional	Integrated
LLaVA-1.5	3.44 (0.72)	3.06 (0.71)	2.96 (0.87)	2.90 (0.97)
ShareGPT4V	3.20 (0.95)	3.36 (0.78)	3.33 (0.92)	3.12 (0.92)
GalleryGPT	2.15 (0.95)	3.29 (0.69)	2.89 (0.87)	2.36 (0.75)
AllCuration (Ours)	**3.93** (0.78)	**4.02** (0.61)	**3.57** (0.85)	**3.56** (1.05)

To examine whether the observed differences among the models were statistically significant, we initially considered applying a one-way analysis of variance (ANOVA). However, as the Integrated category violated the homogeneity of variance assumption (Levene's test, $p = .005$), we adopted the Kruskal–Wallis H-test as a non-parametric alternative across all categories for analytical consistency. The results, summarized in Table 5, reveal statistically significant differences across all evaluation categories ($p < .001$).

5.3 Ablation Study

In this study, we apply SelfCheckGPT during dataset refinement to remove sentences that are

Table 5: Mean ranks of each model across evaluation categories (Kruskal–Wallis H-test)

Model	Cultural	Formal	Emotional	Integrated
LLaVA-1.5	112.49	71.61	88.03	98.25
ShareGPT4V	99.25	94.17	110.02	109.10
GalleryGPT	48.48	90.32	80.55	65.16
AllCuration (Ours)	**141.78**	**145.90**	**123.40**	**129.49**

Table 6: Effect of the SelfCheckGPT threshold on model performance (%). B: BLEU, G: GLEU, M: METEOR, R: ROUGE, BS: BERTScore

Threshold	B	G	M	R	BS
0.1	9.53	18.99	37.93	29.76	90.27
0.2	9.53	18.96	37.63	29.64	90.24
0.3	9.39	18.77	37.15	29.52	90.19
0.4	9.79	19.10	37.83	29.88	**90.29**
0.5	**10.84**	**20.09**	**39.09**	**30.87**	90.28

likely to contain hallucinated information, and conduct an ablation study to analyze the effect of different threshold values.

We vary the threshold from 0.1 to 0.5 and construct separate datasets for each setting. The model is trained under identical settings, and the generated artwork analyses are evaluated using BLEU, GLEU, METEOR, ROUGE, and BERTScore. As shown in Table 6, $\tau = 0.5$ achieves the strongest overall performance across most metrics, while lower thresholds generally lead to reduced performance. This suggests that overly strict filtering may remove not only hallucinated sentences but also informative cultural descriptions, thereby reducing training data diversity. In contrast, a relatively relaxed threshold better balances the factual reliability and data diversity. Therefore, we set $\tau = 0.5$ as the final threshold for our experiments.

5.4 Analysis of SelfCheckGPT Filtering

To analyze the effect of SelfCheckGPT-based refinement, we manually examined cultural-analysis sentences filtered using the threshold $\tau = 0.5$. The filtered sentences were categorized into three semantic reasoning types inspired by Noël Carroll's framework of art criticism (Carroll, 2009): Art-Historical Description (art styles and historical attributes), Cultural Interpretation (cultural and contextual meanings), and Artist Inference (reasoning about the artist's intention or background). Table 7 summarizes the distribution and hallucination score statistics for each category. Although the average hallucination scores are similar across categories, their distributions and qualitative characteristics differ notably.

Table 7: Distribution and hallucination-score statistics of filtered cultural-analysis sentences categorized by semantic reasoning type.

Category	Percentage	Avg score	Std
Historical Description	14.8%	**0.583**	0.167
Cultural Interpretation	40.7%	0.511	0.300
Artist Inference	**44.4%**	0.570	0.259

Artist Inference accounts for the largest proportion of filtered sentences (44.4%) and exhibits a relatively high average hallucination score (0.57). This suggests that interpretations related to an artist's intention or psychology are often not clearly grounded in the artwork itself, making them more speculative and less consistent across generations. In contrast, Art-Historical Description remains relatively stable across generations, while Cultural Interpretation shows broader variability ranging from grounded contextual explanations to more speculative historical or social narratives.

Overall, these findings suggest that SelfCheckGPT-based refinement tends to suppress culturally speculative or weakly grounded interpretations. Although this analysis does not guarantee factual correctness, it provides empirical evidence that the refinement process can help mitigate the potentially unreliable cultural reasoning generated by GPT-4o.

6 Conclusion

Artistic criticism emerges through the interaction of formal, emotional, and cultural elements, and this study jointly considers these dimensions in an integrated manner. To reflect this perspective, we propose **AllCuration**, a fine-tuned MLLM, and construct the ArtCura dataset by extending EmoArt with cultural analysis generated by GPT-4o, and refined using SelfCheckGPT. AllCuration achieved strong performance across various quantitative evaluation metrics and showed meaningful results in both human and statistical analyses. These findings highlight the potential of integrating visual, emotional, and cultural dimensions in multimodal artwork analysis and provide a foundation for future research on multimodal art understanding.

Limitations

Although the proposed framework demonstrates promising results, several limitations remain. Although SelfCheckGPT was used to evaluate the consistency of the generated cultural analyses, this approach does not utilize expert-validated or museum-curated ground-truth descriptions and, therefore, cannot fully guarantee the historical and cultural correctness of the generated interpretations. Future work will incorporate expert-curated references and external knowledge sources to further improve the factual reliability of the cultural analyses.

References

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. https://arxiv.org/abs/2308.12966.

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

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. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7734–7743.

Noël Carroll. 2009. *On criticism*. Routledge.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021. Discourse understanding and factual consistency in abstractive summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 435–447.

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

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

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. Lora: Low-rank adaptation of large language models. In *Proceedings of International Conference on Learning Representations 2022*, pages 1–13.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, and 1 others. 2023. Language is not all you need: aligning perception with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 72096–72109.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Ruixiang Jiang and Chang Wen Chen. 2025. Multimodal llms can reason about aesthetics in zero-shot. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6634–6643.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9332–9346.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of International conference on machine learning 2023*, pages 19730–19742. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fang Liu, Mohan Zhang, Baoying Zheng, Shenglan Cui, Wentao Ma, and Zhixiong Liu. 2023a. Feature fusion via multi-target learning for ancient artwork captioning. *Information Fusion*, 97:101811.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 34892–34916.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. Gleu: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351.

Sebastian Padó and Kerstin Thomas. 2025. Artwork interpretation with vision language models: A case study on emotions and emotion symbols. In *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, pages 1–12.

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

Aashish Anantha Ramakrishnan, Aadarsh Anantha Ramakrishnan, and Dongwon Lee. 2025. Rona: Pragmatically diverse image captioning with coherence relations. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 74–86.

Cheng Zhang, Hongxia Xie, Bin Wen, Songhan Zuo, Ruoxuan Zhang, and Wen-Huang Cheng. 2025. Emoart: A multidimensional dataset for emotion-aware artistic generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12644–12650.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *Proceedings of International Conference on Learning Representations 2020*, pages 1–43.

Chenyue Zhou, Mingxuan Wang, Yanbiao Ma, Chenxu Wu, Wanyi Chen, Zhe Qian, Xinyu Liu, Yiwei Zhang, Junhao Wang, Hengbo Xu, and 1 others. 2025. From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models. https://arxiv.org/abs/2509.25373.