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, resulting in fragmented and shallow interpretations. To address this limitation, an integrated framework is proposed that jointly considers formal, emotional, and cultural dimensions of artworks. We construct the **ArtCura** dataset by extending EmoArt with culturally grounded analyses generated by GPT-4o and refined using SelfCheckGPT to ensure reliability. Based on this dataset, **AllCuration** is developed as a multimodal large language model fine-tuned on a ShareGPT4V backbone to generate comprehensive art critiques. Experimental results demonstrate that the proposed model consistently outperforms existing multimodal baselines across standard evaluation metrics. Human evaluations using Likert-scale assessments show that our model produces more balanced, coherent, and contextually grounded analyses, with strong inter-rater agreement. These results highlight that integrating formal elements, emotional responses, and cultural context is essential for generating coherent and well-structured art critiques. The code is available at https://anonymous.4open.science/r/AllCuration-7682/.

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 (Ramakrishnan et al., 2025; Liu et al., 2023a), visual question answering (Antol et al., 2015), and multimodal reasoning (Zhou et al., 2025; Huang et al., 2023). 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)

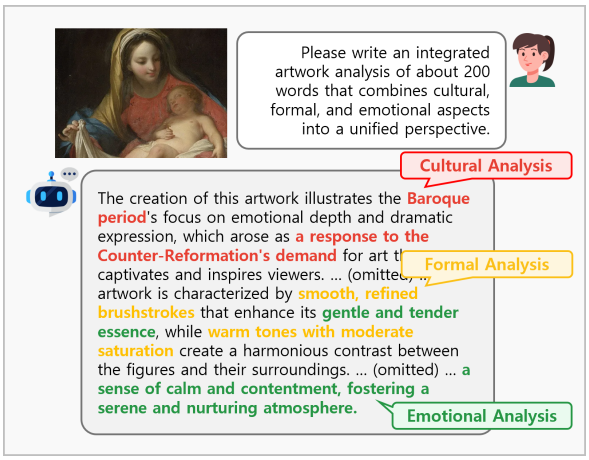


Figure 1: Illustration of AllCuration's integrated art analysis, where cultural(red), formal(yellow), and emotional(green) dimensions are coherently combined into a single, unified analysis.

However, existing approaches to art analysis tend to focus on only a subset of the dimensions involved in art interpretation. For instance, GalleryGPT (Bin et al., 2024) focuses on generating descriptions based on formal elements such as color, composition, and lighting. ArtEmis (Achlioptas et al., 2021) models emotional responses evoked by artworks, capturing the subjective aspects of aesthetic experience. Meanwhile, AQUA (Garcia et al., 2020) incorporates art-historical knowledge and cultural context to support richer interpretations. While these approaches provide valuable insights, most of them treat these aspects in isolation. As a result, frameworks that integrate formal form, emotional expression, and cultural context within a unified multimodal analysis framework remain limited.

This limitation is particularly critical because art

interpretation is inherently a multidimensional process. The meaning of an artwork is not determined solely by its visual form, but emerges from the interaction between perceptual structure, emotional responses, and the cultural context in which the work is created and interpreted. According to Noël Carroll's theory of art criticism(Carroll, 2009), art interpretation is a structured process composed of multiple interconnected stages, including description, contextualization, interpretation, analysis, and evaluation. In this framework, description and analysis address the formal and perceptual aspects of an artwork, contextualization incorporates historical and cultural background, and interpretation and evaluation consider the meaning and emotional effects conveyed by the work. This perspective highlights that meaningful art analysis requires an integrated understanding of visual form, emotional response, and cultural context.

However, existing datasets for multimodal art understanding do not sufficiently support such integrated analysis. For example, datasets such as ArtEmis(Achlioptas et al., 2021) and EmoArt(Zhang et al., 2025) primarily focus on emotional responses and visual attributes of artworks. While these datasets have enabled important progress in modeling affective responses to art, they provide limited support for reasoning about cultural context, such as artistic movements and artist-related information. As a result, it remains challenging to train models capable of generating comprehensive art critiques that integrate formal, emotional, and cultural dimensions.

To address these limitations, 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 cultural, formal, and emotional dimensions of artworks. To improve the reliability of the generated cultural analyses, we further apply the 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.

- We construct **ArtCura** dataset by extending the EmoArt dataset with GPT-4o-generated cultural analyses. To our knowledge, this is one of the first datasets to integrate the cultural, formal, and emotional dimensions of art.

- To enhance the reliability of the cultural analyses, we apply the SelfCheckGPT framework to mitigate hallucinations that may arise during text generation.

- We propose **AllCuration**, a ShareGPT4V-based model fine-tuned on ArtCura, which generates comprehensive art critiques by integrating formal, emotional, and cultural aspects.

2 Related Work

2.1 MLLMs for Art Analysis

Recently, MLLMs have achieved remarkable performance in various vision-language tasks such as image captioning (Ramakrishnan et al., 2025; Liu et al., 2023a) and Visual Question Answering (VQA) (Antol et al., 2015), and have begun to attract attention in art analysis and critique generation.

A representative example is GalleryGPT (Bin et al., 2024), which generates formal analysis of paintings using an instruction dataset of image–text pairs. It demonstrates that leveraging formal visual elements such as color, composition, and lighting can enhance artwork understanding. However, by primarily relying on visual information, it fails to capture cultural context and emotional interpretation, limiting its ability to produce deep and comprehensive analyses.

Meanwhile, some studies highlight that art understanding requires not only visual perception but also art-historical knowledge and contextual information. For example, AQUA (Garcia et al., 2020) distinguishes between visual and knowledge-based questions, emphasizing knowledge-driven reasoning. However, such approaches treat art understanding as decomposed question–answering tasks rather than a unified interpretive process, failing to integrate visual, emotional, and cultural dimensions (Garcia and Vogiatzis, 2018). Therefore, achieving a deeper and more comprehensive understanding of artworks requires an integrated approach that encompasses multiple interpretive dimensions.

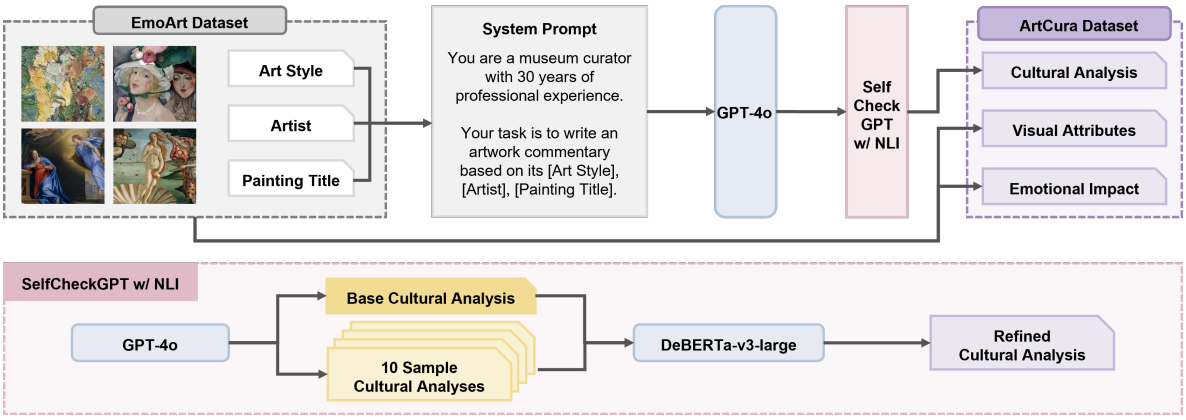


Figure 2: ArtCura dataset construction pipeline

2.2 Datasets for Multimodal Art Understanding

Research on datasets for multimodal art understanding has primarily evolved around the formal elements of artworks and emotional responses. ArtEmis(Achlioptas et al., 2021) is a large-scale dataset constructed to capture emotional reactions to visual artworks, providing images along with emotion labels and corresponding natural language descriptions. Unlike conventional captioning datasets, ArtEmis emphasizes the subjective nature and diversity of emotional interpretations evoked by artworks. EmoArt(Zhang et al., 2025) is a multidimensional dataset that includes a wide range of paintings and provides structured information such as content descriptions, visual attributes, and emotional annotations, including both Valence-Arousal dimensions and categorical emotions, enabling systematic analysis of the relationship between formal visual elements and emotional responses.

However, these datasets primarily focus on visual and emotional dimensions and do not sufficiently support interpretations grounded in cultural context, such as artistic movements or artist-related information. Therefore, there is a need for datasets that integrate formal, emotional, and cultural dimensions, which would provide a critical foundation for more comprehensive and context-aware art analysis.

2.3 Mitigating Hallucination in LLM-Generated Content

Although large language models (LLMs) demonstrate strong text generation capabilities, the issue of hallucination, where models generate factually incorrect or unsupported information, remains a critical challenge as it undermines the reliability and credibility of generated content(Ji et al., 2023; Maynez et al., 2020; Hayashi et al., 2024). Furthermore, widely used automatic evaluation metrics, such as BLEU and ROUGE, have limitations in adequately detecting factual inconsistencies(Kryściński et al., 2020; Gabriel et al., 2021).

To address this issue, SelfCheckGPT(Manakul et al., 2023) has been proposed as a method for assessing the factuality of LLM-generated text without relying on external data or ground-truth annotations. It leverages the principle that information that consistently appears across multiple responses generated through stochastic sampling from the same prompt is more likely to be factual.

Because SelfCheckGPT does not require external knowledge retrieval or annotated evidence, it is well-suited for large-scale text generation settings. In this work, we apply SelfCheckGPT during the dataset construction process to generate and verify cultural background information of artworks, thereby ensuring factual reliability without costly manual verification.

3 Proposed Method

3.1 Dataset Construction

To achieve a comprehensive understanding of artworks, it is necessary to consider not only formal characteristics but also the emotional expressions and cultural context embedded in the work. Visual elements such as color, composition, and brushwork convey external characteristics, but their meanings become clearer when interpreted in relation to the emotional expression embedded in the artwork and its socio-historical context. Therefore, a dataset that integrates formal, emotional, and cultural dimensions is essential for comprehensive artwork analysis

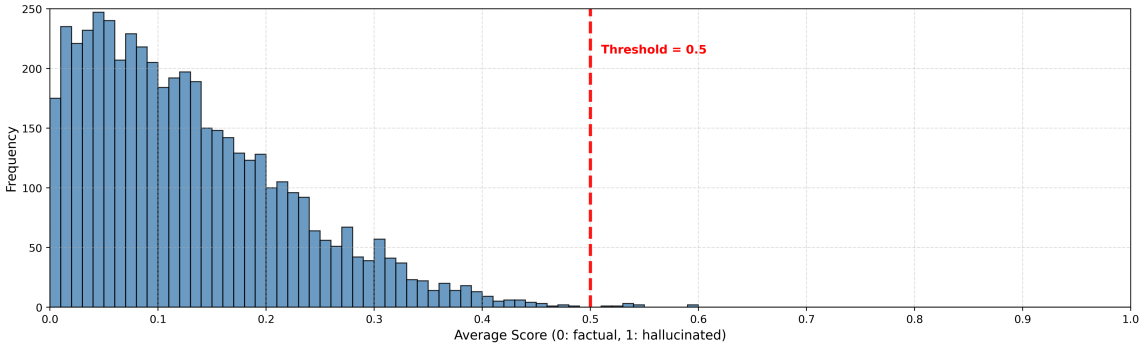


Figure 3: Distribution of SelfCheck NLI average scores

However, existing multimodal datasets primarily focus on formal descriptions or emotion annotations, often lacking sufficient cultural information. To address this limitation, we construct the ArtCura dataset, which extends the EmoArt(Zhang et al., 2025) dataset by incorporating 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 provided 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 within specific social and historical contexts.

To address this gap, we generate cultural analyses for 5,132 artworks using the GPT-4o large language model. We design a rule-based prompt to produce historically grounded cultural analyses, avoiding formal descriptions and speculative interpretations. The prompt instructs the model to act as an experienced museum curator and base its analysis on verifiable art-historical knowledge.

Because this process relies on external knowledge sources, such as artists' biographies and socio-historical contexts, we exclude 324 artworks with incomplete or ambiguous metadata. For the remaining artworks, GPT-4o generates culturally grounded analyses, explaining how each artwork was shaped by its cultural and historical context.

3.2 Dataset Refinement

Large language models, such as GPT-4o, can provide cultural analyses, but they may also produce plausible yet factually incorrect information, known as hallucinations. 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. We used a threshold of $\tau = 0.5$, determined through the ablation study described in Section 4.3, and removed sentences that exceeded this threshold, resulting in only nine sampels being eliminated. After this refinement process, the final ArtCura dataset comprises 4,799 images with culturally and historically grounded analyses.

To analyze the impact of this filtering strategy, Figure 3 presents the distribution of the average SelfCheckGPT-NLI scores computed over all the cultural analyses in the dataset. The distribution is

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 influencing 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 the 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, warm).
Integrated	Q10	The description provides a well-balanced integration of cultural, formal and emotional perspectives.

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.

We split the 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 variants per sample through instruction paraphrasing, resulting in 11,517 training instances. The test set was not augmented and consists of 960 samples to ensure a fair evaluation.

3.3 AllCuration: Supervised Fine-Tuning

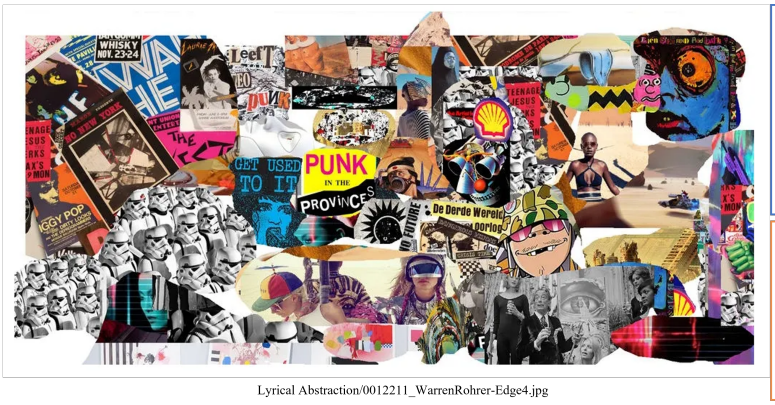
To verify whether the ArtCura dataset can enhance the capability of MLLMs in generating comprehensive art critiques, we developed AllCuration. We implemented supervised fine-tuning (SFT) with LoRA (Hu et al., 2022) on the ShareGPT4V (Chen et al., 2024) backbone using ArtCura to achieve parameter-efficient optimization. This approach allows the model to effectively adapt to the art critique domain without requiring full parameter updates, thereby reducing computational cost while maintaining performance. The SFT process utilized instruction–answer pairs that aligned each painting with its corresponding formal, emotional and cultural analyses. We fine-tuned the model on an RTX 6000 GPU with a learning rate of 2e-5 and batch size of 16. For the LoRA module, we set the LoRA rank and alpha values to 128 and 256, respectively.

Unlike existing vision–language models that primarily focus on formal features, AllCuration aims to generate integrated artwork analyses by jointly considering formal, emotional, and cultural dimensions. Through supervised fine-tuning on the ArtCura dataset, the model learns to associate each painting with analyses across these three dimensions, enabling the generation of context-aware and comprehensive art critiques.

4 Experiments

4.1 Quantitative Analysis

To evaluate the effectiveness of the proposed model, AllCuration, we conducted experiments to evaluate the model's ability to generate integrated artwork analyses across formal, emotional, and cultural dimensions. A total of 960 paintings from the SelfCheckGPT-validated dataset were used as the test set to assess model accuracy. We compared AllCuration against several open-source multimodal baselines, including LLaVA-1.5 (Liu et al., 2023b), ShareGPT4V (Chen et al., 2024), Qwen2.5-VL (Bai et al., 2025), and GalleryGPT (Bin et al., 2024). All models were evaluated using the same prompt: "*Please write a 200-word artwork analysis that explicitly situates the piece within its historical and cultural context and then discuss its formal and emotional aspects.*"



Lyrical Abstraction/0012211_WarrenRohrer-Edge4.jpg

AllCuration (Ours)

Emerging in the late 20th century, Neo-Pop Art represents a reaction to the swift pace of technological progress and the cultural changes of the post-Soviet era. This artistic movement captures the essence of consumerism and the commodification of art, reflecting the complexities of identity and memory in a society undergoing significant transformation. The artist's background in graphic design and commercial art further shapes the work, emphasizing the convergence of high and low culture, as well as the influence of mass media on contemporary life. The artwork is characterized by smooth, digital brushstrokes that blend seamlessly, creating a lively and chaotic composition that lacks a central focus. The color palette is dominated by bright, saturated hues, with high contrast that enhances the visual experience. The sharp, defined lines contribute to a graphic, digital aesthetic, while the limited application of light and shadow relies on color contrast to create depth. This combination of elements results in an engaging and vibrant visual experience that captures the viewer's attention. The vibrant colors and dynamic composition evoke excitement and energy, while the chaotic arrangement and graphic quality introduce a sense of intrigue and unpredictability.

GalleryGPT

The artwork under scrutiny is a vibrant and chaotic mosaic of magazine clippings, each piece a testament to the cultural and social landscape of the 1980s. The visual characteristics of the piece are a testament to the artist's ability to manipulate and assemble these fragments into a cohesive whole. The use of bold colors, particularly the stark contrast of black and white, adds a sense of drama and intensity to the piece. The emotional resonance of the artwork is equally compelling, with the eclectic mix of images evoking a sense of nostalgia and unease. The magazine clippings, ranging from mainstream to underground, create a visual symphony of popular culture, reflecting the fragmented and complex nature of society during that decade. The philosophical influences of the art movement are evident in the artist's use of collage as a tool to explore themes of identity, consumption, and the constructed nature of reality. The artwork is a powerful visual commentary on the cultural and social context of the 1980s, encapsulating the chaos and creativity of the era.

LLaVA-1.5

The image is a collage of various posters and images, including Star Wars characters, a man wearing a tie, and a woman wearing a hat. The collage is a mix of different styles and themes, showcasing a diverse range of visual elements. The decade in which this collage was created is likely the 1970s, as it reflects the cultural and social environment of that time. The art movement associated with this collage is likely punk rock, as evidenced by the presence of Star Wars characters and the overall eclectic nature of the collage. The formal aspects of the collage include the use of different sizes, colors, and orientations of the images, which create a sense of visual chaos and dynamism. The emotional aspect of the collage is its ability to evoke a sense of nostalgia and a connection to the cultural and social environment of the 1970s. The collage also serves as a representation of the counterculture movement, which challenged societal norms and values during that time. The use of Star Wars characters, for example, was a way for the punk rock movement to express their rebellion against the mainstream culture. Overall, the collage is a powerful representation of the cultural and social environment of the 1970s, as well as the emotions and philosophical influences of the punk rock movement.

ShareGPT4V

The image is a vibrant collage that captures the essence of punk rock culture in the 1970s and 1980s. The collage is composed of various posters, flyers, and photographs, each contributing to the overall theme. The dominant colors are black, white, and pink, with occasional splashes of yellow and blue adding contrast and visual interest.The collage is a visual representation of the punk rock movement, which emerged as a counterculture to mainstream society in the 1970s and 1980s. The posters and flyers feature slogans and images that reflect the movement's rebellious spirit, such as "Get used to it" and "Punk in the provinces". These texts and images are scattered throughout the collage, creating a sense of chaos and disorder that is characteristic of the punk rock aesthetic.The collage also includes photographs of people, likely punk rock musicians and fans, further emphasizing the movement's social and cultural significance. The presence of these images adds a personal and emotional dimension to the collage, inviting viewers to connect with the movement's values and ideals.In terms of formal aspects, the collage's arrangement of images and text creates a dynamic visual experience. The overlapping and juxtapositioning of elements add depth and complexity to the piece, drawing the viewer's eye across the various elements and encouraging them to explore the collage in detail.Overall, the collage is a powerful representation of punk rock culture, capturing its rebellious spirit, social significance, and emotional impact. It invites viewers to engage with the movement's values and ideals, and to appreciate the visual and emotional impact of this influential cultural phenomenon.

Figure 4: Qualitative comparison of model-generated analyses for artwork

To ensure consistency, all generations were produced at a temperature of 0.2, which minimized randomness during the evaluation. 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., 2019). Artwork analysis can be formulated as an image-grounded description generation task that incorporates visual, cultural, and emotional aspects. Accordingly, image captioning metrics are suitable for evaluating the quality of the generated descriptions and their semantic consistency with reference texts. In particular, BERTScore is effective in capturing cultural and emotional meanings that may be expressed in diverse ways (Zhang et al., 2019).

4.2 Qualitative Analyses

To evaluate the ability of large language models (LLMs) to generate insightful and interpretable art criticism, we conducted a human evaluation comparing five multimodal models: LLaVA-1.5, ShareGPT4V, Qwen2.5-VL-7B, GalleryGPT, and our proposed model. Ten participants were randomly divided into two groups, each evaluating model-generated descriptions for five artworks. Participants compared 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 evaluation reliability.

As summarized in Table 1, 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 (Carroll, 2009). This categorical structure enables a systematic and fine-grained analysis of multiple complementary dimensions of art criticism rather than relying on a single overall quality score. Specifically, the Cultural category (Q1–Q3) assessed whether the descriptions appropriately reflected the historical background, artistic movement, and broader societal or temporal context of the artwork. The Formal category (Q4–Q6) evaluated how effectively the descriptions articulated formal visual properties, including composition, balance, spatial structure, color usage, lighting, and surface-level features such as brushstrokes and line quality. The Emotional category (Q7–Q9) focused on the extent to which the descriptions conveyed emotions, emotional tone, and affective atmosphere evoked by the artwork through appropriate emotional language. Finally, the Integrated category (Q10) measured the overall coherence and balance with which the cultural, formal, and emotional perspectives were combined into a unified critical interpretation.

To assess the internal consistency of the survey items, we computed Cronbach's α 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. Although the Formal category shows relatively lower α values, it remains within an acceptable range for exploratory evaluation. Overall, these results support the proposed survey design as a reliable instrument for analyzing qualitative differences in model-generated descriptions.

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

Evaluation CategoryAspect	Cronbach's α
Cultural (Q1–Q3)	0.731
Formal (Q4–Q6)	0.649
Emotional (Q7–Q9)	0.742

4.3 Ablation Study

In this study, we apply SelfCheckGPT during dataset refinement to remove sentences that are likely to contain hallucinated information. We set the SelfCheckGPT-NLI threshold to $\tau = 0.5$ in our main setting 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 training settings, and the generated artwork analyses are evaluated using BLEU, GLEU, METEOR, ROUGE, and BERTScore. As shown in Table 3, $\tau = 0.5$ achieves the strongest overall performance across most evaluation 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 the diversity of the training data. In contrast, a relatively relaxed threshold better balances factual reliability and data diversity. Therefore, we set $\tau = 0.5$ as the final threshold in our experiments.

5 Results

We evaluated the performance of AllCuration in generating artwork analyses and compared it with several multimodal baselines. Table 4 presents the quantitative comparison results based on standard image captioning metrics. GalleryGPT shows relatively strong textual fluency, as reflected in its high GLEU score, while Qwen2.5-VL achieves strong performance on METEOR but more moderate results on other metrics. In contrast, AllCuration consistently achieves strong performance

Table 3: 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

Table 4: 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
AllCuration (Ours)	**10.84**	**20.09**	**39.09**	**30.87**	**90.28**	Passed

across all evaluation metrics, indicating more balanced overall results. This suggests that AllCuration effectively captures and integrates formal, emotional, and cultural dimensions in artwork analysis. These results demonstrate that the combination of a SelfCheckGPT-validated dataset and supervised fine-tuning on the ShareGPT4V backbone contributes to improved generation of comprehensive and integrated artwork analyses.

To further assess the factual reliability of the generated descriptions, we conducted a systematic manual hallucination inspection of the top-10 samples per model, ranked by BERTScore. Three independent evaluators examined each instruction–response pair, explicitly assessing the presence of three predefined categories of factual errors: (1) temporal errors, referring to incorrect historical periods or anachronistic references; (2) art-movement errors, including misclassification or conflation of artistic styles and movements; and (3) artwork title errors, encompassing incorrect, imprecise, or fabricated names of artworks or artists. 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 minor discrepancies.

The inspection results revealed distinct error patterns across the models. GalleryGPT and LLaVA-1.5 exhibit temporal errors. In contrast,

Table 5: Descriptive analysis of human evaluation scores

Model	Cultural	Formal	Emotional	Overall
LLaVA-1.5	3.44 (0.72)	3.06 (0.71)	2.96 (0.87)	3.13
ShareGPT4V	3.20 (0.95)	3.36 (0.78)	**3.33** (0.92)	3.28
GalleryGPT	2.15 (0.95)	3.29 (0.69)	2.89 (0.87)	2.74
AllCuration (Ours)	**3.91** (0.82)	**3.98** (0.68)	2.87 (0.99)	**3.54**

Table 6: Mean ranks of each model across evaluation categories (Kruskal–Wallis H-test)

Model	Cultural	Formal	Emotional	Overall
LLaVA-1.5	376.35	291.64	343.40	1128.68
ShareGPT4V	341.07	349.75	**413.28**	1221.07
GalleryGPT	185.72	338.31	325.25	912.18
AllCuration (Ours)	**458.48**	**465.14**	324.49	**1383.87**

ShareGPT4V demonstrated a higher prevalence of art movement-related errors. Qwen2.5-VL, however, exhibited a consistently high incidence across all three error categories, with temporal inaccuracies, art-movement misclassifications, and artwork title errors appearing in a relatively balanced and recurrent manner. Quantitatively, hallucinations were identified in four out of ten inspected cases for GalleryGPT, six cases for LLaVA-1.5, five cases for ShareGPT4V, and seven cases for Qwen2.5-VL. Notably, the hallucinations produced by Qwen2.5-VL were not only more frequent but also more severe, often involving multiple factual errors across different categories within a single answer. Given both the high number of errors and their wide range, Qwen2.5-VL was considered unreliable for detailed artwork analysis and was therefore excluded from subsequent human evaluation and statistical analysis.

The human evaluation results are summarized in Table 5. AllCuration achieved the highest mean score in the Formal category, indicating a clear strength in effectively capturing the structural and compositional characteristics of the artworks. Although it shows relatively lower scores in the Emotional category, AllCuration demonstrates competitive performance in terms of the overall average and the Integrated category. Notably, its performance in the Integrated category suggests that the model effectively generates analyses that comprehensively reflect formal, emotional, and cultural dimensions.

To examine whether the observed differences among the models were statistically significant, we initially considered applying a one-way ANOVA. However, this approach was deemed inappropriate because of a violation of the homogeneity of variance assumption, as indicated by Levene's test ($p < .001$). Consequently, we adopted the Kruskal–Wallis H-test as a non-parametric alternative that does not require equal variances across groups. The results, summarized in Table 6, reveal statistically significant differences across all evaluation categories for the remaining models.

6 Conclusion

We propose AllCuration, a fine-tuned MLLM that integrates formal, cultural, and emotional dimensions. To support this, we build the ArtCura dataset by extending EmoArt with culturally grounded annotations generated by GPT-4o and refined using SelfCheckGPT. Experimental results show that AllCuration achieves strong performance across evaluations. Both human and statistical analyses highlight its strengths in formal and structural analysis, as well as its overall interpretive depth. Although its emotional expressiveness is relatively limited, the model maintains balanced interpretations and strong alignment with human perspectives.

More importantly, this study goes beyond prior approaches that focus on individual dimensions and presents an integrated framework for artwork analysis that jointly incorporates formal, emotional, and cultural aspects. This unified approach reflects that artistic meaning arises from the interaction of multiple elements, and human evaluation shows that such integrated analyses are perceived as more comprehensive and satisfying.

Limitations

While 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 fully guarantee factual accuracy. The evaluation relies on consistency verification rather than comparisons with expert-validated or museum-curated ground-truth descriptions, which limits the ability to fully confirm the historical and cultural correctness of the generated interpretations. Nevertheless, the framework is extensible, and future work will incorporate expert-curated references and external knowledge sources to further improve the factual reliability of 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.

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 and George Vogiatzis. 2018. How to read paintings: semantic art understanding with multimodal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

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, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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. *Advances in Neural Information Processing Systems*, 36:72096–72109.

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.

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 *International conference on machine learning*, pages 19730–19742. PMLR.

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

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. *Advances in neural information processing systems*, 36:34892–34916.

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. *arXiv preprint arXiv:2005.00661*.

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.

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. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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. *arXiv preprint arXiv:2509.25373*.