

Leveraging Vision–Language Models for Semantic Interpretation of Historical Paintings: A Case Study on Religiousness

Yuya Kanazawa, Midori Tanaka, Hiroshi Kera, Takahiko Horiuchi; Chiba University; Chiba, Japan

Abstract

Historical paintings reflect the social, cultural, and religious contexts of their time. With the emergence of vision–language models (VLMs), it has become possible to generate textual interpretations from images; however, it remains unclear what information these models rely on and how their outputs should be evaluated. This study examines the characteristics and validity of VLM-generated interpretations through two experiments. First, an art style classification task using the Pandora dataset shows that VLMs tend to group paintings into historically related styles, although strict distinctions are not always achieved. Second, focusing on religious paintings, we evaluate the agreement between generated interpretations and museum descriptions using BERTScore under three conditions: image only, image with metadata, and metadata only. Results indicate that metadata improves scores, while visual input has limited impact. Moreover, evaluation outcomes depend strongly on the content of reference texts. These findings suggest that VLM-based interpretation relies more on linguistic context than visual information and highlight limitations of using museum descriptions as evaluation references.

Introduction

Historical paintings are not only valuable as works of art but also constitute important cultural materials that reflect the social, cultural, and religious contexts of the periods in which they were created. The humanistic knowledge expressed in paintings—such as the ideas, values, and social transformations of each era—has been studied across multiple disciplines, centered on art history and including cultural anthropology, archaeology, social history, psychology, and religious studies [1]. In contrast, within the field of engineering, research has developed around the measurement and analysis of paintings, largely motivated by the preservation of cultural heritage. Material analysis techniques such as X-ray imaging, infrared imaging, and spectroscopic analysis have been established to investigate pigments and canvas structures in paintings. Through these methods, the technical characteristics of specific historical periods have gradually been clarified [2][3]. However, these studies have primarily focused on the material and physical aspects of paintings, and the relationship between such physical properties and the intellectual or cultural contexts of artworks has not been sufficiently discussed. As a result, a substantial gap still exists between the physical information obtained through engineering analyses and the interpretative knowledge developed in the humanities. Against this background, the long-term goal of this research is to clarify the relationship between the physical features of paintings and their associated humanistic knowledge. Specifically, we aim to construct a model that takes physical features as input and humanistic knowledge as output, and to analyze and visualize the internal representations of the model to reveal relationships between physical properties and knowledge. Nevertheless, realizing such a model requires the

preparation of a large and diverse set of humanistic knowledge texts associated with each painting as training data. Because systematically annotating such knowledge manually is constrained by practical costs, it is necessary to develop methods for automatically collecting humanistic knowledge.

In this study, we focus on Vision–Language Models (VLMs) as a possible approach for automatically collecting such knowledge. VLMs are machine learning models that process both images and natural language, enabling them to generate textual descriptions from visual content and respond to language-based queries. Recently, models combining large language models with visual encoders have demonstrated powerful performance in various vision–language tasks, including image caption generation and Visual Question Answering (VQA). However, it remains unclear to what extent VLMs can accurately understand and express specialized knowledge embedded in artworks, such as historical background, religious context, and cultural meaning. To address this challenge, Hayashi et al. evaluated the ability of VLMs to generate artwork descriptions by comparing model outputs with reference texts derived from Wikipedia articles using similarity metrics such as BERTScore [4]. Their results showed that providing artwork titles significantly improved the evaluation scores, whereas performance decreased when only images were used. These findings suggest that VLMs rely more heavily on linguistic cues than on visual information when generating descriptions of artworks. However, the reference texts used in their study were primarily general-purpose descriptions from Wikipedia and may not reflect domain-specific or scholarly interpretations.

To address this limitation, this study focuses on religious paintings and uses official museum descriptions as reference texts, which are expected to better reflect domain-specific interpretations. By comparing these texts with explanations generated by VLMs, we investigate whether VLMs can generate knowledge representations grounded in specialized contexts rather than merely producing general descriptive knowledge. In particular, we evaluate whether VLMs can serve as a reliable source for collecting humanistic knowledge.

VLM Art Style Classification

This experiment serves as a preliminary analysis to examine whether VLMs can extract visual stylistic features from paintings.

Procedure

Dataset: Pandora

In this study, the Pandora (Paintings Dataset for Recognizing the Art Movement) dataset proposed by Florea et al. was used to investigate whether vision–language models (VLMs) can identify art styles from the visual characteristics of paintings [5]. The Pandora dataset consists of 7,724 painting images categorized into 12 art movements. The target art movements include Old Greek pottery, Iconoclasm, High Renaissance, Baroque, Rococo, Romanticism, Realism, Impressionism, Fauvism, Cubism,

Surrealism and Abstract expressionism. Example images from the dataset are shown in Figure 1. As summarized in Table 1, the dataset covers a wide range of art movements spanning a broad historical timeline, from ancient art to modern and contemporary styles.



Figure 1. Example images from the Pandora dataset.

Table 1. Overview of the Pandora dataset.

Art Movement	No.	Historical period
Old Greek pottery	350	Antiquity
Iconoclasm	665	Middle age
High renaissance	812	1490 – 1527
Baroque	960	1590 – 1725
Rococo	844	1650 – 1850
Romanticism	874	1770 – 1850
Realism	307	1848 – present
Impressionism	984	1860 – 1925
Fauvism	426	1900 – 1950
Cubism	920	1900 – present
Surrealism	242	1900 – present
Abstract expressionism	340	1920 – present

Task Setup

In this study, LLaVA-1.5 was used as the vision–language model (VLM) [6]. To investigate the inherent capability of the model in interpreting paintings, we directly used the open-source model llava-hf/llava-1.5-7b-hf available on Hugging Face without additional training or fine-tuning. We formulated an art style classification task in which the model was required to select one art movement from a set of 12 candidate styles corresponding to the categories defined in the Pandora dataset. The prompt used for the model is shown in Figure 2. In this prompt, the model was instructed to output only one style name from the predefined list, while the generation of explanations or additional text was explicitly

prohibited. This constraint was introduced to eliminate ambiguous free-form responses and to ensure that the classification results could be clearly aggregated. As a result, the model’s predictions for each style could be objectively analyzed.

Please look at the image and answer with only one of the following styles. No explanation.

[old Greek pottery, iconoclasm, high renaissance, baroque, rococo, romanticism, impressionism, realism, cubism, fauvism, abstract-expressionism, surrealism]

Answer:

Figure 2. Prompt for the art style classification task.

Results and Discussion

To examine the classification tendencies of the model, art style classification was performed on 12 art movements from the Pandora dataset using LLaVA-1.5. The results are summarized as a confusion matrix in Table 2, where the horizontal axis represents ground-truth labels and the vertical axis represents predicted labels.

As shown in Table 2, correct predictions are concentrated along the diagonal for several styles. For example, 827 out of 984 Impressionism images and 725 out of 960 Baroque images were correctly classified. However, some styles were rarely selected as predictions; no images were classified as Rococo, and only one image was classified as Realism. This indicates that the model tends to favor certain styles over others regardless of the input.

To further analyze these tendencies, the confusion matrix was normalized column-wise and visualized as a heatmap (Table 3), where each value represents the relative proportion for each ground-truth label. The results show that several styles retain relatively strong diagonal components, suggesting that the model can identify some art movements to a certain extent. In addition, styles such as Baroque, Romanticism, Impressionism, and Abstract Expressionism exhibit high values not only on the diagonal but also in neighboring categories.

This pattern indicates that misclassifications are not random but tend to occur among historically or stylistically related styles. These results suggest that although LLaVA-1.5 does not strictly reproduce art-historical classifications, it captures coarse stylistic relationships in paintings. This tendency is consistent with previous studies [7], which report that neural networks organize visually similar styles into nearby regions in their internal representation space.

Quantitative Evaluation of Religious Paintings

Procedure

Target Artworks

To evaluate VLM-generated interpretations using museum descriptions as reference texts, artwork data and corresponding descriptions were obtained through the public API provided by the Cleveland Museum of Art (CMA). For the selection of target artworks, paintings depicting the Annunciation were chosen as representative cases with clearly defined religious themes and well-established interpretations in art history. The search criteria included artworks whose title or description contained the keyword “Annunciation.” Based on this query, six paintings were extracted from the CMA collection. Figure 3 presents the six selected paintings along with their corresponding CMA artwork IDs. The retrieved dataset includes the artwork ID, title, artist name, creation

Table 2. Confusion matrix of art style classification results.

	Gre	Icon	Ren	Bar	Roc	Rom	Real	Impr	Fauv	Cub	Sur	Abs
Gre	226	3	0	7	0	0	0	2	0	0	6	104
Icon	2	195	0	468	0	0	0	0	0	0	0	0
Ren	0	0	121	682	0	1	0	3	0	0	4	1
Bar	2	0	18	725	0	72	0	98	35	0	9	1
Roc	1	0	8	523	0	188	0	105	10	0	9	0
Rom	0	0	17	250	0	331	0	188	22	0	63	3
Real	0	0	0	17	0	94	0	124	31	0	29	12
Impr	4	0	0	11	0	101	0	827	0	0	40	1
Fauv	0	0	0	1	0	8	1	289	3	0	47	49
Cub	0	0	0	10	0	5	0	203	8	17	342	335
Sur	0	0	0	3	0	3	0	5	0	0	169	61
Abs	0	0	0	1	0	0	0	72	0	0	47	220

Table 3. Heatmap of the normalized confusion matrix for art style classification results.

	Gre	Icon	Ren	Bar	Roc	Rom	Real	Impr	Fauv	Cub	Sur	Abs
Gre	96.1	1.51	0	0.25	0	0	0	0.1	0	0	0.78	13.2
Icon	0.85	98.4	0	17.3	0	0	0	0	0	0	0	0
Ren	0	0	73.7	25.2	0	0.12	0	0.15	0	0	0.52	0.12
Bar	0.85	0	10.9	26.8	0	8.96	0	5.11	32.1	0	1.17	0.12
Roc	0.42	0	4.87	19.3	0	23.4	0	5.48	9.17	0	1.17	0
Rom	0	0	10.3	9.26	0	41.2	0	9.81	20.1	0	8.23	0.38
Real	0	0	0	0.63	0	11.7	0	6.47	28.4	0	3.79	1.52
Impr	1.7	0	0	0.4	0	12.5	0	43.1	0	0	5.22	0.12
Fauv	0	0	0	0.03	0	0.99	100	15	2.75	0	6.14	6.22
Cub	0	0	0	0.37	0	0.62	0	10.5	7.33	100	44.7	42.5
Sur	0	0	0	0.11	0	0.37	0	0.26	0	0	22	7.75
Abs	0	0	0	0.03	0	0	0	3.75	0	0	6.14	27.9

date, museum description, and the associated image data. To manage these elements consistently, a JSON-based data structure was designed in this study. For each artwork, the structure integrates the basic metadata, the museum description used as the reference text, and the explanations generated by the VLM.



Figure 3. Annunciation paintings.

Experimental Conditions

We hypothesize that VLM-generated interpretations rely more heavily on linguistic metadata than on visual input when dealing with abstract concepts such as religious meaning. In this experiment, three experimental conditions—base, meta, and noimage—were defined to examine how differences in input information provided to the VLM affect the generation of interpretations related to religious context. For each condition, the prompt and the availability

of image input were varied. In the base condition, the painting image was provided as input to the model, and the prompt “Explain the religious significance of this artwork.” was used. The purpose of this condition was to examine the extent to which the model can verbalize the religious significance of the artwork based solely on visual information. In the meta condition, in addition to the image input used in the base condition, metadata describing the artworks—such as the title, creator, and creation date—were explicitly included in the prompt. Specifically, the prompt followed the format: “Explain the religious significance of this artwork. For context: Title = ‘...’, Creator = ‘...’, Date = ‘...’.” This condition provides both visual information and basic contextual information about the artwork. In the noimage condition, the painting image was not provided as input, and the same prompt used in the meta condition was applied. In this case, only the metadata were given to the model. This condition allows us to examine the extent to which the model can generate religious interpretations using textual information alone, without visual input. By comparing these three conditions, the experimental design enables us to separately evaluate the contributions of visual information and metadata in the generation of religious interpretations. In addition, four Vision–Language Models were employed in the experiments: LLaVA-1.5, LLaVA-1.6, LLaVA-NeXT, and Qwen2.5-VL [8]. These models were selected to compare the behavior of different VLM architectures under the same experimental conditions.

Evaluation Method

In this study, BERTScore was used as a quantitative metric to evaluate the extent to which the religious interpretations generated by the VLM are consistent with museum descriptions [9].

BERTScore, proposed by Zhang et al., is an automatic evaluation metric for comparing textual similarity and has been reported to show a higher correlation with human judgments than conventional n-gram-based metrics in natural language generation tasks. A key characteristic of BERTScore is that it measures semantic similarity between texts based on contextualized embeddings rather than relying on surface-level matches such as word overlap. Specifically, it computes similarity scores between tokens using contextual representations derived from pretrained language models. In this study, the museum descriptions provided by the Cleveland Museum of Art were used as reference texts, while the explanations generated by the VLM were treated as candidate texts. BERTScore was then calculated between the reference and candidate texts to quantitatively evaluate the degree of agreement in religious interpretation. However, the suitability of museum descriptions as reference texts is further examined in the Results section.

Results and Discussion

Comparison of Scores Across Models and Conditions

BERTScore was calculated for six religious paintings using four VLMs—LLaVA-1.5, LLaVA-1.6, LLaVA-NeXT, and Qwen2.5-VL—under three experimental conditions: base, meta, and noimage. Table 4 shows the average scores computed across the six paintings. Across all models, the F1 scores were concentrated within a narrow range of approximately 0.83–0.84, and no substantial differences were observed among the VLMs. When comparing the average scores across experimental conditions, the base condition achieved an average score of 0.834, the meta condition 0.838, and the noimage condition 0.840. These results indicate that slightly higher scores were obtained when metadata were provided together with the image or when metadata alone were used, compared with the condition relying solely on visual input. This result suggests that the language model alone may already contain strong prior knowledge about religious themes, reducing the contribution of visual input.

Table 4. Average BERTScore by Model and Condition.

Model	base	meta	noimage
LLaVA1.5	0.830	0.837	0.843
LLaVA1.6	0.839	0.837	0.841
LLaVA-NeXT	0.835	0.841	0.837
Qwen2.5-VL	0.833	0.838	0.840

T-test Analysis and Discussion

To examine whether the differences in BERTScore across conditions were statistically significant rather than caused by random variation, a paired t-test was conducted. A paired t-test is a statistical method used to determine whether the difference between the means of two conditions is significant while accounting for the variance of paired observations. In this study, BERTScores obtained from six paintings and four VLMs under each condition were used to perform paired t-tests between experimental conditions. The results are presented in Table 5.

Table 5. Results of the Paired t-test.

base-meta	base-noimage	meta-noimage
0.047	0.022	0.449

The paired t-test results showed that statistically significant differences at the 5% significance level were observed between the

base and meta conditions, as well as between the base and noimage conditions. In contrast, no significant difference was found between the meta and noimage conditions. These results indicate that providing metadata, such as the artwork title, creator name, and creation date, contributes to improving the agreement between the generated religious interpretations and the museum descriptions. In other words, in the task of interpreting religious context using VLMs, conditions that include explicit linguistic context information tend to achieve higher semantic similarity with museum descriptions than the condition relying solely on visual input. On the other hand, the absence of a significant difference between the meta and noimage conditions suggests that, within the scope of the religious interpretation task and evaluation metric used in this study, the inclusion of visual information does not necessarily contribute to improving similarity with the reference descriptions.

Limitations of Museum Descriptions

Since no substantial differences were observed among the models, the BERTScores were averaged across models to compare the six artworks under the three experimental conditions. The results are summarized in Table 6.

Table 6. Average BERTScore by Painting and Condition.

id	Base	meta	noimage
97176	0.821	0.825	0.831
122338	0.848	0.859	0.861
127572	0.851	0.858	0.860
130981	0.826	0.830	0.827
131820	0.828	0.821	0.821
515375	0.831	0.838	0.839

The analysis revealed that certain artworks consistently produced either higher or lower scores regardless of the input condition. In other words, the variation in BERTScore tended to cluster by artwork rather than by experimental condition. To investigate the cause of this tendency, two representative examples with a clear score difference were examined under the same model and condition (LLaVA-1.6, base condition): a low-score case (Artwork ID 97176, BERTScore = 0.819) and a high-score case (Artwork ID 122338, BERTScore = 0.855). As shown in Figure 4, the museum description for Artwork ID 97176 primarily discusses issues of author attribution. Although the term Annunciation appears, the text does not mainly aim to interpret or explain the religious theme of the painting. In contrast, the output generated by LLaVA-1.6 provides a general explanation of religious iconography and the relationship between visual elements and religious symbolism. Because the reference text contains little explicit religious interpretation, the semantic overlap between the generated text and the museum description remains limited, resulting in a relatively low BERTScore. By contrast, as shown in Figure 5, the museum description for the high-score example (Artwork ID 122338) focuses directly on the theme of the Annunciation and explains the roles of key figures such as the Virgin Mary and the Archangel Gabriel. The generated output from LLaVA-1.6 also correctly captures the theme using religious vocabulary such as Annunciation, Mary, and Gabriel, leading to a higher semantic overlap with the reference text and consequently a higher BERTScore. In this experiment, the prompt instructed the model to explain the religious significance of the painting, meaning that the generated texts aim to articulate religious context explicitly. Therefore, if the reference museum description does not explicitly describe religious meaning, the generated interpretation may still be

appropriate while receiving a relatively low score in terms of BERTScore. These observations suggest that the score differences observed in this study are influenced less by the VLM’s ability to understand religious context and more by how explicitly the museum descriptions themselves express religious interpretation. Museum descriptions are generally intended to provide comprehensive information about artworks, including provenance, stylistic background, and historical context, rather than systematically explaining religious symbolism alone. Consequently, using museum descriptions directly as reference texts may have inherent limitations when evaluating the generation of specialized humanistic interpretations embedded in historical paintings.

Conclusion

This study examined how vision–language models (VLMs) generate and interpret knowledge from paintings, with a focus on the roles of visual and linguistic information. The results indicate that VLMs can capture coarse stylistic characteristics from visual input, but the generation of religious interpretations depends more strongly on linguistic contextual information, such as metadata.

Furthermore, the evaluation results were found to be highly sensitive to the choice of reference texts, suggesting limitations in using museum descriptions as a ground truth for assessing interpretation validity. These findings highlight a fundamental challenge in evaluating humanistic knowledge generated by VLMs.

Future work should focus on designing more appropriate reference texts that better reflect domain-specific knowledge, as well as developing evaluation methods that go beyond similarity-based metrics. In addition, domain-specific adaptation of VLMs may improve the integration of visual features and humanistic knowledge.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 25H01930.

References

- [1] A. Hauser, “Social history of art”, Routledge, 1951.
- [2] S. Tominaga, H. Ujike and T. Horiuchi, “Surface Reconstruction of Oil Paintings for Digital Archiving”, Proc. IEEE SSIAP, pp.173–176, 2010.
- [3] S. Tominaga, Y. Imai, R. Saito, and T. Horiuchi, “Extraction of Artists' Color Features of Art Paintings and its Application to Color Image Correction,” Color and Imaging Conference, vol. 21, 2013.
- [4] K. Hayashi, Y. Sakai, H. Kamigaito, K. Hayashi, T. Watanabe, “Towards Artwork Explanation in Large-scale Vision Language Models,” Proceedings of the 2024 Conference of the Association for Computational Linguistics: Short Papers, pp. 705–729, 2024.



Museum description	Prominent art historians still debate the identity of the artist and the original function of this painting. First attributed to the Neapolitan artist Colantonio (active 1440-1460) and also to the Flemish painter Berthélemy d'Eyck--known as the Master of the Aix Annunciation (active 1440-1470)--the painting is now more likely considered a product of the Provençal school. Undoubtedly it exemplifies the stylistic influences of the Netherlands, Provence, and Italy in the 1400s. The subject may be either a physician or an Old Testament prophet from an altarpiece.
LLaVA1.6	The artwork you've provided appears to be a portrait of a religious figure, likely a saint or a religious leader, given the attire and the setting. The figure is depicted with a red cloak, which is often associated with religious figures in art, symbolizing their status and the importance of their role. The use of a red cloak is a common artistic convention to denote saints or holy figures, as red is often associated with martyrdom and the blood of Christ. The figure's pose and the setting suggest a solemn and contemplative mood, which is fitting for a religious figure. The background

Figure 4. Museum description and LLaVA-1.6 generated text for Artwork ID 97176 (base condition).



Museum description	The scene of the Annunciation to Mary that she will give birth to Christ is depicted here as a contemporary scene that may have been familiar to any viewer at the end of the 1400s. Mary is portrayed as an ordinary but wealthy woman who is interrupted by the Archangel Gabriel as she reads a book of hours. This type of book was enormously popular in the late Middle Ages as a prayer and devotional book, particularly among the wealthy, literate nobility, and city aristocracy.
LLaVA1.6	The artwork you've provided appears to be a representation of the Annunciation, a common theme in Christian art. The Annunciation is the story of the angel Gabriel announcing to the Virgin Mary that she would conceive and bear a son, Jesus Christ. In the image, we see two figures: one is a woman dressed in a blue robe, which is a common color for the Virgin Mary in Christian art, symbolizing purity and divinity. The other figure is an angel, identifiable by the wings, which is a common depiction of angels in Christian iconography. The angel is holding

Figure 5. Museum description and LLaVA-1.6 generated text for Artwork ID 122338 (base condition).

- [5] C. Florea, R. Condorovici, C. Vertan, R. Butnaru, L. Florea and R. Vrânceanu, "Pandora: Description of a Painting Database for Art Movement Recognition with Baselines and Perspectives", 2016 24th European Signal Processing Conference (EUSIPCO), pp.918–922, 2016.
- [6] H. Liu, C. Li, Y. Li and Y. Lee, "Improved Baselines with Visual Instruction Tuning", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.26296–26306, 2024.
- [7] A. Elgammal, B. Liu, D. Kim, M. Elhoseiny and M. Mazzone, "The Shape of Art History in the Eyes of the Machine", Proceedings of the AAAI Conference on Artificial Intelligence, Vol.32, No.1, 2018.
- [8] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song et al., "Qwen2.5-VL Technical Report", arXiv:2502.13923, 2025.
- [9] T. Zhang, V. Kishore, F. Wu, K. Weinberger, Y. Artzi, "BERTScore: Evaluating Text Generation with BERT", arXiv:1904.09675, 2020.

Author Biography

Yuya Kanazawa is an undergraduate student at Chiba University, Japan, studying computer vision and multimodal learning, with a focus on vision–language models for the analysis of historical artworks. His research explores the evaluation of semantic interpretations generated by VLMs using museum descriptions and quantitative metrics.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



- **SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •**
- **INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •**

www.electronicimaging.org

