Art Appreciation of Western Classical Art Painting

Ziying Li
Department of Computer Science
Rice University
ziying.li@rice.edu

Yuruo Gong
Department of Computer Science
Rice University
yuruo.gong@rice.edu

Xiaowei Xu
Department of Computer Science
Rice University
xiaowei.xu@rice.edu

Abstract

Art appreciation is a task that demands substantial creativity, knowledge accumulation, and possesses a strong subjectivity, making it a long-standing challenge in machine learning tasks. In other words, finding a balance between powerful associative capabilities and established facts has always been a worthwhile research topic in art appreciation. To enhance the accuracy of art appreciation, this paper integrates Convolutional Neural Networks (CNNs) and Vision Transformers (BEiT) using the WikiArt Dataset[1] *to train models capable of recognizing painting styles. Subsequently, this new model combined with Retrieval-Augmented Generation (RAG) as a supplementary tool for LLaVA in the appreciation of paintings to improve the authenticity of generated sentences. To balance associative capabilities with established facts, we employ several evaluation methods including Rouge-1, Meteor Score, Key Aspect Rating, and Human Ratings. These methods have been demonstrated to significantly enhance the comprehensiveness and reliability of art appreciation.*

1. Introduction

Art painting appreciation by computers is complex. It involves recognizing painting elements and translating these into descriptive text. This often uses multimodal models that handle both images and text. The common method is to extract features from paintings using CNNs, and then use an adapter to make these features compatible with text models. Text models then generate language based on this information and additional context like the author or the painting's title [2, 1, 3].

In our study, we train models with the WikiArt Dataset to identify painting styles. We also use Retrieval-Augmented Generation (RAG) to get information about the painting, which helps the language model create better descriptions. Our advanced model, Input-Optimized LLaVa, uses this information to improve its descriptions (figure1).

Moreover, we utilize the LLaVA model, which is skilled at processing both images and text. It uses GPT-4 to create multimodal instructions, which then help fine-tune the LLaVa model for specific tasks like art appreciation. This lets the model convert the intricate details of art into text descriptions, providing deeper insights into the paintings.

By refining the LLaVA model, we aim to enhance the way computers appreciate art and relate the artwork's essence to the audience.

2. Related Work

Early work on art appreciation includes [2], which utilizes a nearest neighbor approach based on image features img_k and text features com_k & att_k to learn the relationship between images and texts, thereby achieving an understanding of art appreciation. Intermediate work includes first generating masked sentences, providing a fill-in-the-blank question based on the image topic, then searching online for relevant answers and filling them in [1]. Later, instead of using masked sentences, prompts were used. In other words, the association between text and images is understood through the CLIP model, and then multimodal information along with a prompt is passed to a large language model to generate art appreciation. Early work lacked creativity, while later work could not discern the accuracy of appreciation. Therefore, we choose to first equip the image model with greater classification ability on the details of art paintings, then provide established facts about the paintings as keywords to find a balance between creativity and accu-

[1] https://paperswithcode.com/dataset/wikiart

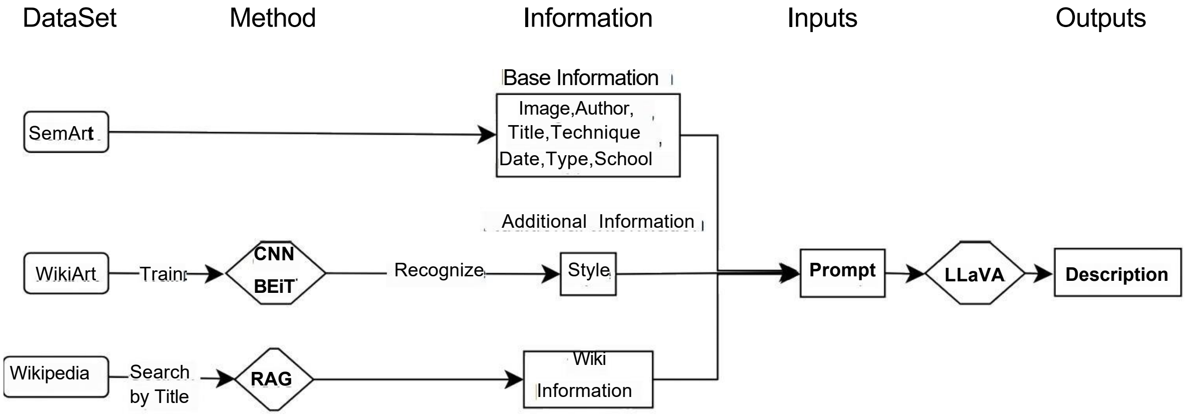


Figure 1. Inputs and Outputs for the LLaVA

Model	ResNet50	BEiT	ResNet50&BEiT
Accuracy	55%	62%	71%

Table 1. Comparison of style classification models

racy.

3. Model

3.1. Data Collection

For each image, the style, the wikipedia information, along with the SemArt Dataset's text-input and image are fed into a multimodal model.

3.1.1 Style Classifier Model

We have chosen BEiT (Vision Transformer) and CNN (Convolutional Neural Network) to recognize the styles of art paintings because BEiT possesses a self-attention mechanism capable of modeling relationships across the entire image. In contrast, CNNs focus on local perceptions and excel in capturing features such as textures and edges of images. For tasks that require understanding of both the global context and local brushstrokes, such as analyzing the styles of art paintings, BEiT and CNN exhibit good complementarity (see Table 1).

In our model, we employed the WikiArt training set to train the style recognition model. The WikiArt dataset categorizes paintings from the 15th century to the present into 27 styles. Since style is an abstract concept, training models to recognize painting styles has always been a challenging topic. Here, we assigned weights of 0.57 and 0.43 to BEiT and CNN, respectively, and achieved satisfactory results.

3.1.2 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a powerful technique that addresses the limitations of large language models by leveraging an external, authoritative knowledge base, such as Wikipedia. This approach significantly enhances the model's contextual understanding and helps strike a balance between creative appreciation and factual accuracy in the domain of art appreciation.

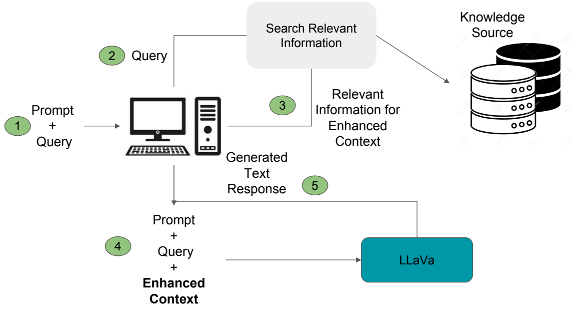

Search Relevant Information
Knowledge Source
2 Query
1 Prompt + Query
3 Relevant Information for Enhanced Context
Generated Text Response 5
4 Prompt + Query + Enhanced Context
LLaVa

Figure 2. Workflow of RAG

One of the primary challenges in using large language models for art appreciation is their potential to generate illusory or inconsistent outputs. To mitigate this issue, we employ RAG, which augments the model's context by retrieving relevant information from Wikipedia based on the painting's title. This external, authoritative knowledge serves as a prompt, guiding the language model to generate more factually grounded and contextually relevant appreciations. By integrating this retrieved information, RAG effectively balances the model's creative interpretation of the artwork with the necessary factual foundation, thus reducing the risk of generating nonsensical or misleading content.

The retrieved Wikipedia introduction is input alongside the style information, enabling the model to incorporate both the painting's historical and cultural background and its artistic style into the generated appreciation. By integrating this contextual information, RAG effectively balances the model's creative interpretation of the artwork with the necessary factual foundation, reducing the risk of generating nonsensical or misleading content. Moreover, RAG helps ensure that the generated appreciations are not only creative but also factually accurate. By grounding the model's output in the authoritative information provided by Wikipedia, we maintain a higher level of reliability and consistency in the generated content. This is particularly important in the domain of art appreciation, where subjective interpretations must be balanced with objective facts and historical context.

3.2. LLaVa

LLaVA is a multimodal model that combines a vision encoder and a large language model (LLM) for general-purpose visual and language understanding [4]. The key innovation is the use of language-only GPT-4 to generate multimodal language-image instruction-following data, which is then used to instruction-tune the LLaVA model. This approach enables LLaVA to effectively understand and reason about image content according to instructions, making it well-suited for our art appreciation task.

One of the main reasons we chose LLaVA is its ability to capture and convey the details of paintings to the language model. In classic art paintings, the intricate details often reflect the emotions and intentions of the artist. LLaVA's visual instruction tuning approach allows it to "translate" these details into a text sequence that the language model can understand and reason about. This is particularly important for our task, as it enables the model to provide insightful and nuanced descriptions of the paintings.

4. Evaluation Method

We conducted a comprehensive evaluation of the Input-Optimized LLaVa model. To assess its performance in terms of "Association" and "Reality", we employed various evaluation methods, including Key Aspect score, ROUGE-1, METEOR score and human ratings. These tools enabled us to precisely measure the quality of the model's output. In addition, we compare the enhanced LLaVa-1.5 model's performance with high-performance multimodal models, including the original LLaVa-1.5 and MiniCPM-V.

4.1. ROUGE-1

ROUGE-1 compares the overlap of unigrams (individual words) between the machine-generated text and the reference texts. It measures the number of overlapping words to determine how well the generated text captures the content of the reference texts. There are two primary components of the ROUGE-1 score:

$$R = \frac{\text{number of overlapping unigrams}}{\text{total number of unigrams in reference summary}}$$

$$P = \frac{\text{number of overlapping unigrams}}{\text{total number of unigrams in generated summary}}$$

By ensuring a high degree of keyword overlap between the text and high-quality reference text, ROUGE-1 helps to verify the authenticity and reliability of the generated text information.

4.2. METEOR Score

Metric for Evaluation of Translation with Explicit ORdering is an automatic metric for evaluating the quality of text generated by machine translation systems. METEOR can recognize synonyms as valid translations, not just exact word matches. By evaluating the fluency and naturalness of translation, METEOR helps ensure that the generated text is closely aligned with human natural expression in language use, enhancing its realism.

4.3. Human Rating

To evaluate the effectiveness of the generated art descriptions, a human evaluation is conducted. This assessment gauges how well the descriptions created by the model reflect the paintings' artistic qualities and factual accuracy. We recruited 35 classmates to participate in our human ratings. We show a 3 generated appreciations to each annotator, together with the image and the original SemArt comment. We ask annotators to rate each description according to the metrics below (higher is better):

Understandability: Rated on a scale from 1 to 5, this metric assesses how easily the description can be understood by a layperson.

Relevance: Also rated from 1 to 5, this measures the pertinence of the description to the painting it describes.

Accuracy: This assesses the factual correctness of the description, rated from 1 to 5.

Informativeness: Measures the depth and richness of the description, indicating how well it captures the essence and various aspects of the painting, such as style, emotion, and technique. Split richness into smaller parts, 1 if it includes this content, 0 otherwise.

5. Experiments and Results

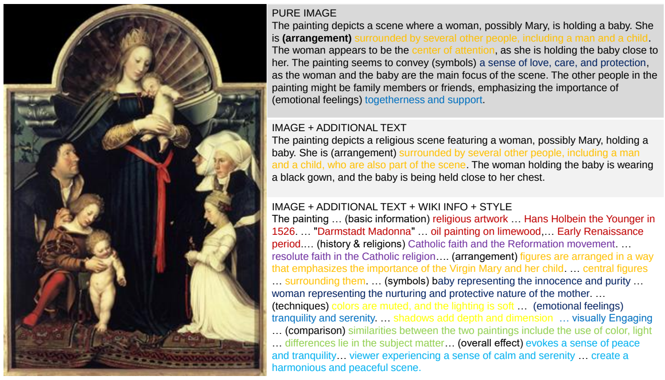
5.1. Outputs Comparison

To better illustrate the experimental results, we present a comparative analysis of the generated art descriptions for a well-known painting. Figure 3 showcases the outputs from three different models: the original model, which only takes the input picture; the original model augmented with additional information about the painting; and our Input-Optimized Model.

As evident from the figure, the output generated by our Input-Optimized Model offers the richest and most comprehensive art appreciation. The description not only captures the visual elements and style of the painting but also incorporates relevant historical and cultural context. Moreover, the generated text maintains a high level of factual accuracy, demonstrating the model's ability to balance creative interpretation with objective information.

In contrast, the original model, LLaVa-1.5, which relies solely on the input picture, generates a description that, while capturing some visual aspects, lacks depth and contextual understanding. The augmented original model, which incorporates additional information about the painting, shows improvement in terms of context but still falls short in terms of richness and coherence compared to our Input-Optimized Model.

These results underscore the effectiveness of our approach in leveraging both visual and textual information to generate comprehensive and accurate art appreciations. By optimizing the input through style classification and retrieval-augmented generation, and by employing the LLaVa architecture, our model achieves a balance between the richness of appreciation and factual accuracy, ultimately providing a more engaging and informative experience for the viewer.



PURE IMAGE
The painting depicts a scene where a woman, possibly Mary, is holding a baby. She is **(arrangement)** surrounded by several other people, including a man and a child. The woman appears to be the center of attention, as she is holding the baby close to her. The painting seems to convey (symbols) a sense of love, care, and protection, as the woman and the baby are the main focus of the scene. The other people in the painting might be family members or friends, emphasizing the importance of (emotional feelings) togetherness and support.

IMAGE + ADDITIONAL TEXT
The painting depicts a religious scene featuring a woman, possibly Mary, holding a baby. She is (arrangement) surrounded by several other people, including a man and a child, who are also part of the scene. The woman holding the baby is wearing a black gown, and the baby is being held close to her chest.

IMAGE + ADDITIONAL TEXT + WIKI INFO + STYLE
The painting … (basic information) religious artwork … Hans Holbein the Younger in 1526. … "Darmstadt Madonna" … oil painting on limewood,… Early Renaissance period.… (history & religions) Catholic faith and the Reformation movement. … resolute faith in the Catholic religion…. (arrangement) figures are arranged in a way that emphasizes the importance of the Virgin Mary and her child. … central figures … surrounding them. … (symbols) baby representing the innocence and purity … woman representing the nurturing and protective nature of the mother. … (techniques) colors are muted, and the lighting is soft … (emotional feelings) tranquility and serenity. … shadows add depth and dimension … visually Engaging … (comparison) similarities between the two paintings include the use of color, light … differences lie in the subject matter… (overall effect) evokes a sense of peace and tranquility… viewer experiencing a sense of calm and serenity … create a harmonious and peaceful scene.

Figure 3. Darmstadt Madonna, Hans Holbein the Younger, 1526

5.2. Evaluation results

Model	P	R	METEOR
MiniCPM-V	**19.8**	17.64	23.9
LLaVa-1.5	10.2	7.1	15.8
Ours	15.9	**32.7**	**24.2**

Table 2. Comparison of models on ROUGE-1 and METEOR.

Table 2 presents the comparison of different models on ROUGE-1 and METERO. P in ROUGE-1 measures the proportion of the generated summary's unigrams (individual words) that are also found in the reference summary. From the result, MiniCPM-V has the highest P score, and our model is only is only inferior(19.8 vs. 15.9 vs. 10.2). Due to the fact that the appreciate our model generated is much longer than the appreciate generated by the other two models and the reference, although our overlap with the reference is high, our own proportion will still decrease as a result.

R in ROUGE-1 measures the proportion of the reference summary's unigrams that are captured by the generated summary. We can see that our model excels in R score with a score of 32.7, which is significantly higher than both other models(32.7 vs. 17.64 vs. 7.1). During the evaluation, our model's R score has reached a maximum value of 65. This result indicates that our Input-Optimized LLaVa Model is particularly good at capturing a wide range of relevant information.

Model	U	R	A	I
SemArt	2.8	2.5	2.3	0.35
LLaVa-1.5	3.2	2.9	2.7	0.41
Ours	**4.5**	**4.3**	**4.1**	**0.83**

Table 3. Model comparison on Human Rating

Table 3 presents the comparison of different models based on the human rating metrics. Our Input-Optimized LLaVa Model outperforms both the SemArt dataset and the original LLaVa-1.5 model across all metrics.

The human evaluation results demonstrate that our Input-Optimized LLaVa Model outperforms both the SemArt dataset and the original LLaVa-1.5 model across all metrics. Our model achieves significantly higher scores in understandability (4.5 vs. 2.8 and 3.2), relevance (4.3 vs. 2.5 and 2.9), accuracy (4.1 vs. 2.3 and 2.7), and informativeness (0.83 vs. 0.35 and 0.41). These results indicate that our model generates descriptions that are more easily comprehensible, closely related to the paintings, factually correct, and richer in content compared to existing datasets and models.

These human evaluation results validate the effectiveness of our Input-Optimized LLaVa Model in generating art descriptions that are more understandable, relevant, accurate, and informative compared to existing datasets and models. The combination of enhanced style classification, retrieval-augmented generation, and the use of LLaVa architecture enables our model to produce high-quality art appreciations that balance creativity and factual accuracy.

References

[1] J. Fumanal-Idocin, J. Andreu-Perez, O. Cordon, H. Hagras, and H. Bustince. Artxai: Explainable artificial intelligence curates deep representation learning for artistic images using fuzzy techniques. *arXiv preprint arXiv:2308.15284*, 2023.

[2] N. Garcia and G. Vogiatzis. Semantic art understanding with multi-modal retrieval. *arXiv preprint arXiv:1810.09617*, 2018.

[3] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image

encoders and large language models. *arXiv preprint arXiv:YourArXivNumberHere*, 2023.

[4] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. NeurIPS 2023 Oral; project page: https://arxiv.org/abs/2304.08485.