ArtRAG: Structured Context Retrieval-Augmented Framework for Artwork Explanation Generation

Shuai Wang[1], Ivona Najdenkoska[1] Hongyi Zhu[1], Stevan Rudinac[1], Monika Kackovic[1], Nachoem Wijnberg[1], and Marcel Worring[1]

University of Amsterdam

1 Introduction

Generating detailed and rich-context explanations for paintings is a challenging problem at the intersection of computer vision and natural language processing. Unlike general image captioning, describing artwork requires not only recognizing visual elements but also understanding their artistic, cultural, and historical contexts. These descriptions often span multiple dimensions, for example including content (what is depicted), form (how it is depicted, e.g., style and technique), and context (why and when it was created, including historical and cultural influences [1]).

Current methods for incorporating Art-contextual information primarily rely on static attributes (e.g., title, artist, and medium) and fixed relationships (e.g., "painted by," "belongs to movement"). While such structured graphs have proven effective for tasks like painting classification [2] and explanation generation [4], they lack the flexibility to capture dynamic and multi-faceted contextual information. Many paintings—particularly those with deep cultural or historical significance—require a more expansive range of contextual data, including historical events, religious movements, and societal changes, to fully explain their richness and meaning.

For instance, understanding Hans Holbein the Younger's "The Ambassadors" as shown in Fig. 1 needs to go beyond attributes and requires knowledge of multiple interconnected factors: Artistic Context: Holbein's role in the Northern Renaissance and his mastery of detailed realism and symbolism; Thematic Context: The memento mori (anamorphic skull) symbolizes mortality and the transient nature of human accomplishments; Historical Context: The painting reflects Renaissance humanism and scientific discovery, while also alluding to religious tensions during the Protestant Reformation. Such intricate interconnections are beyond the representational capacity of art knowledge graphs built from attributes by existing methods, which lack mechanisms to dynamically understand and integrate contextual information across these dimensions. Consequently, existing methods struggle to generate comprehensive multi-topic explanations that fully reflect the depth of these works.

To address this gap, we propose a novel framework ArtRAG, that automatically extracts structured information from art-related text corpora to construct

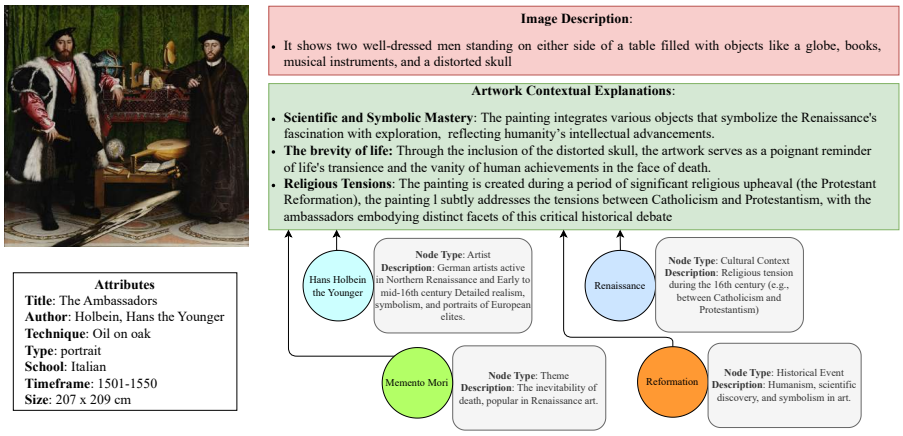

Fig. 1. An example illustrating the interconnected contextual factors necessary for generating a comprehensive description of The Ambassadors by Hans Holbein the Younger. Explaining this painting requires not only information about the visual painting itself and its attributes but also a broader understanding across multiple perspectives, including Artistic Context, Historical and Cultural Context, and Thematic Context.

a rich, context-aware knowledge graph. Our approach captures detailed explanations of the relationships between graph nodes (e.g., artists, historical events, and artistic themes) and edges (e.g., "influenced by," "represents"). By leveraging this structured graph in a retrieval-augmented generation (RAG) framework, our method retrieves and organizes relevant context, including historical details, cultural influences, and stylistic characteristics, associated with a given artwork. This structured retrieval enables our model to generate descriptions that are not only visually grounded but also contextually rich and coherent across multiple perspectives.

To evaluate our method, we utilize the SemART [3] and Artpedia [5] datasets, which include detailed textual annotations for paintings. These annotations cover both visual (visual subjects and objects), and contextual (historical, cultural, and biographical influences) dimensions, providing a robust benchmark for assessing the quality and interpretative depth of generated descriptions. We demonstrated that incorporating structured contextual knowledge through RAG significantly enhances the quality of generated descriptions compared to traditional retrieval-augmented approaches. Overall, our work bridges the gap between visual recognition and contextual understanding in artwork description, offering a robust framework for integrating structured information into generative models.

References

1. Belton, R., of Critical, F., Studies, C.: Art History: A Preliminary Handbook. University of British Columbia (1996), https://books.google.nl/books?id=JHakzQEACAAJ
2. Efthymiou, A., Rudinac, S., Kackovic, M., Worring, M., Wijnberg, N.: Graph neural networks for knowledge enhanced visual representation of paintings. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 3710–3719. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475586, https://doi.org/10.1145/3474085.3475586
3. Garcia, N., Vogiatzis, G.: How to read paintings: Semantic art understanding with multi-modal retrieval (2018)
4. Jiang, Y., Ehinger, K.A., Lau, J.H.: Kale: An artwork image captioning system augmented with heterogeneous graph. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. International Joint Conferences on Artificial Intelligence Organization (2024)
5. Stefanini, M., Cornia, M., Baraldi, L., Corsini, M., Cucchiara, R.: Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In: Image Analysis and Processing – ICIAP 2019 (2019)