Benchmarking Multimodal Large Language Models in Zero-shot and Few-shot Scenarios: Preliminary Results on Studying Christian Iconography

Gianmarco Spinaci[1,2], Lukas Klic[2], and Giovanni Colavizza[1,3]

[1] Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
[2] Villa i Tatti, The Harvard University Center for Italian Renaissance Studies, Firenze, Italy
[3] Department of Communication, University of Copenhagen, Copenhagen, Denmark

Abstract

This study evaluates the capabilities of multimodal large language models (LLMs) in the task of single-label classification of Christian iconography, focusing on their performance in zero-shot and few-shot settings across curated datasets. The goal was to assess whether general-purpose models, such as *GPT-4o* and *Gemini 2.5*, can interpret the Iconography, typically addressed by supervised classifiers, and evaluate their performance. Two research questions guided the analysis: **(RQ1)** How do multimodal LLMs perform on image classification of Christian saints? And **(RQ2)**, how does performance vary when enriching input with contextual information or few-shot exemplars?

We conducted a benchmarking study using three datasets supporting Iconclass natively: ArtDL, ICONCLASS, and Wikidata, filtered to include the top 10 most frequent classes. Models were tested under three conditions: (1) classification using class labels, (2) classification with Iconclass descriptions, and (3) few-shot learning with five exemplars. Results were compared against ResNet50 baselines fine-tuned on the same datasets.

The findings show that *Gemini-2.5 Pro* and *GPT-4o* outperformed the ResNet50 baselines across the three configurations reaching peaks of **90.45%** and **88.20%** in ArtDL, respectively. Accuracy dropped significantly on the Wikidata dataset, suggesting model sensitivity to image size and metadata alignment. Enriching prompts with class descriptions generally improved zero-shot performance, while few-shot learning produced lower results, with only occasional and minimal increments in accuracy.

We conclude that general-purpose multimodal LLMs are capable of classification in visually complex cultural heritage domains, even without specific fine-tuning. However, their performance is influenced by dataset consistency and the design of the prompts. These results support the application of LLMs as metadata curation tools in digital humanities workflows, suggesting future research on prompt optimization and the expansion of the study to other classification strategies and models.

Keywords: Multimodal Models, Large Language Models, Image Classification, Iconography

1 Introduction

For over two decades, the GLAM sector (galleries, libraries, archives, and museums) has undergone an extensive mass digitization process, resulting in a vast amount of digital archives containing a diverse array of artworks, photographs, and documents [9]. This rapid growth is transforming the analogical Cultural Heritage into a body of machine-readable knowledge, defining a critical

Gianmarco Spinaci, Lukas Klic, and Giovanni Colavizza. "Benchmarking Multimodal Large Language Models in Zero-shot and Few-shot Scenarios: Preliminary Results on Studying Christian Iconography." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1197–1209. https://doi.org/10.63744/oxWtm5MhhwBH.

mass of images and their metadata and serving as input to research in Artificial Intelligence (AI) and Computer Vision [20].

Computer Vision is a part of AI and Computer Science that enables computers to analyze, interpret, and make decisions based on image characteristics. At its core, it extracts and analyzes visual patterns, with applications that span the fields of medicine, robotics, and cultural heritage [3; 4], with main tasks that include image classification [5; 11; 32], object detection [8; 10], and semantic segmentation [17]. Among these, **image classification** stands out for its applicability to the semantically and visually complex objects in the Cultural Heritage field. For this task, the previously mentioned datasets are well-suited for image classification, as they also contain metadata, usually created by domain experts, that describes the general content of the images.

An influential area of study supported with Image classification is Iconography, being that *"branch of the history of art which concerns itself with the subject matter or meaning of works of art, as opposed to their form"* [22]. Iconography helps to understand the representations and themes expressed in images by identifying symbols, subjects, and motifs in paintings. Iconclass [6] is a formal tool that can be utilized to support this area of study, offering a thesaurus that catalogs subjects and objects in artworks, including elements of historical, religious, and architectural significance. Thanks to this tool, image classification can be used to understand the overarching theme represented in an artwork, beyond surface-level object detection, without focusing on individual elements. Iconclass has been adopted as the backbone of several datasets, including ArtDL [19] and the Iconclass AI Test Set [23]. Another important dataset is IICONGRAPH [27], which includes metadata for images from Wikidata and ArCo [2], modeled after the ICON Ontology [28], expanding the granularity to the context of Iconology [22], and allowing cross-analyses of images based on the themes they represent.

These datasets have become more significant and widely utilized due to the substantial surge in the field of Computer Vision, where technical advancements enabled the use of Convolutional Neural Networks (CNNs) [21] and, more recently architectures centered on Vision Transformers (ViTs) [7], enabling models to capture global dependencies in images rather than focusing on local features. To this family belong *CLIP* [25] and *SigLIP* [33]. They both align images and text in a shared embedding space through contrastive learning, excelling in zero-shot image classification. Beyond this, ViTs support multimodal Large Language Models (LLMs), including OpenAI *GPT* [24], Google *Gemini* [31], *Claude* [1], *LLaVa* [16], and *Mistral* [12]. General-purpose systems capable of interpreting texts and images together, connecting complex visual and linguistic information. These models can be directly interfaced with full-text queries and analyze pictures, pushing the boundaries of automatic study as they can be adapted to a wide range of tasks without requiring retraining from scratch.

Over the last few years, several initiatives have been launched to leverage multimodal models and achieve state-of-the-art results in image classification on datasets such as Iconclass, that has been subject to several studies [14; 26; 29]. Another example is fine-grained food image recognition through vision-language models [13], leveraging *CLIP* and image descriptions generated with *MiniGPT-4* over data from two different datasets. Another study explores the use of off-the-shelf multimodal models, including *CLIP*, *SigLIP*, and *BLIP-2* [15], in a zero-shot environment for classifying historical photographs in the Estonian Ajapaik archive [18]. The authors found these models to be underperforming in comparison to a fine-tuned supervised baseline for ambiguous or culturally specific categories, such as "viewpoint elevation". For the case of image classification of Christian saints, the literature contains applications of CNNs to classify Christian iconography in artworks. For Example, in the paper introducing *ArtDL* [19], the authors trained a *ResNet50* model, achieving an accuracy of *84.44%* in identifying the depicted saints.

These experiments demonstrate a growing interest in using multimodal models for general-purpose image classification tasks. Many advancements in the state-of-the-art have been achieved

by implementing these models, and most importantly, by enhancing the accuracy scores of these models through the addition of meaningful context to the prompts [13]. Despite this interest, the literature has not yet benchmarked LLMs specifically for classification in Christian iconography. This gap in the literature highlights the novelty of our preliminary investigation, which aims to address the following research questions: **(RQ1)** How do LLMs perform on the classification of Christian saints? **(RQ2)** How do the results change with the progressive enrichment of contextual data, such as adding more descriptive data or tagged exemplars?

2 Methodology

In this analysis, we designed a study to evaluate the classification performance of diverse LLM models on Christian Iconography. Specifically, we aimed to (RQ1) compare different LLM versions in image classification and (RQ2) assess the impact of progressively enriching the input data with descriptions and few-shot exemplars. This section outlines the procedure design, including dataset preparation, model selection, and the benchmarking.

2.1 Datasets

We investigated the images belonging to three collections, which we selected for their native support of Iconclass classes. One is *ArtDL* [19] and the other two are subsets of the *ICONCLASS Test AI set* [23] and *Wikidata*, of which we only include images representing the top ten most frequent classes of Christian Saints.

2.1.1 ArtDL

The *ArtDL* dataset comprises over 42,000 images of Christian religious paintings. Each image is annotated with Iconclass codes, covering 10 key figures of Christian iconography, such as the Virgin Mary, Saint Francis of Assisi, and Saint Sebastian. The test set, published along with the paper, contains **1,864 images**. Each is mapped to a single iconographic label and serves as the starting point for our classification experiments presented in this study.

2.1.2 ICONCLASS AI Test Set

The *ICONCLASS* [23] dataset originates from the official AI Test Set. The original collection comprises approximately 87,500 images representing a wide range of iconographic subjects. For this study, we conducted a filtering and curation process explicitly tailored to the single-label classification of Christian saints. Our refinement began by selecting all images annotated with Iconclass codes, starting with *"11F"* (The Virgin Mary), *"11H"* (male saints), and *"11HH"* (female saints). We then performed a frequency analysis to identify the 10 most common saint classes. We applied systematic controls by removing all images with multiple saints, resulting in a dataset comprising **863 images**.

2.1.3 Wikidata

To expand our benchmark beyond curated archives, we compiled a dataset of religious artworks using the *Wikidata SPARQL endpoint*. We designed a SPARQL query to detect paintings[1] with a valid image URL and associated with Iconclass codes representing Christian saints and the Virgin Mary (using the same codes as the *ICONCLASS* dataset). After filtering out multi-label entries, failed downloads, and duplicates, we retained **718 images** of paintings. Images were retrieved in their original size, preserving native aspect ratios.

[1] In Wikidata are instances of type *"wd:Q3305213"*

2.1.4 Cross-Dataset Similarity Analysis

To assess dataset independence and uncover overlaps that could bias evaluation results, we conducted a cross-dataset similarity analysis using a robust block-based image hashing method inspired by digital image forensics [30]. This approach identifies near-duplicate images across datasets, even those that have been transformed, such as cropping or mirroring. This method is particularly suitable for comparing art datasets, where visual duplicates may originate from different digitization pipelines or cataloging standards, resulting in varying representations of the same artwork. We defined a duplicate as any pair of images across datasets with a Hamming distance of 8 or less, identifying **36 cross-dataset duplicate** pairs using the robust hash, primarily between ArtDL and Wikidata (Figure 1).

2.2 Models

The study compares multimodal LLMs to highlight the differences in paradigms when processing single-labeled images of paintings depicting Christian Saints. The inclusion of these models reflects differences in the scale of parameters, input resolution, and the processing and presentation of visual and textual information.

We assess unified multimodal models, such as OpenAI *GPT-4o (snapshot 2024-08-06)*, Google *Gemini 2.5 Pro (preview-05-0)* classification smaller counterparts *GPT-4o-mini (snapshot 2024-07-18)* and *Gemini 2.5 Flash (preview-05-20)*. Flash and Mini are lightweight versions designed for efficiency and reduced cost while maintaining core multimodal capabilities. These models integrate both modalities into a single processing stream, enabling image understanding and contextual visual reasoning through prompt-based classification with natural language. They are particularly effective for visual studies, description generation, and interpretive tasks that require deeper semantic integration across modalities.

2.3 Benchmarking design

For building the benchmarking, we employed three distinct test configurations applied across models and datasets. This framework assesses model performance under varying conditions of knowledge availability.

Test 1 is a zero-shot classification task with labeled names. The models classify images using only class label names (e.g., “St. Paul” or “Mary Magdalene”) without additional contextual information. This configuration tests the model’s ability to leverage pre-trained knowledge connections between visual features and semantic concepts.

Test 2 is a zero-shot classification task with label descriptions. The Models receive detailed iconographic descriptions for each class, retrieved from Iconclass descriptions (e.g., “The penitent harlot Mary Magdalene; possible attributes: book (or scroll), crown …”). This approach evaluates how effectively models can utilize rich textual descriptions to guide visual classification.

Test 3 is a few-shot Learning classification task. During inference, LLMs are provided with an arbitrary number of five example images, along with their corresponding class labels. The images have been chosen to represent the five among the less performing classes from the first test. This configuration assesses the models’ capacity for rapid adaptation and in-context learning from limited examples.

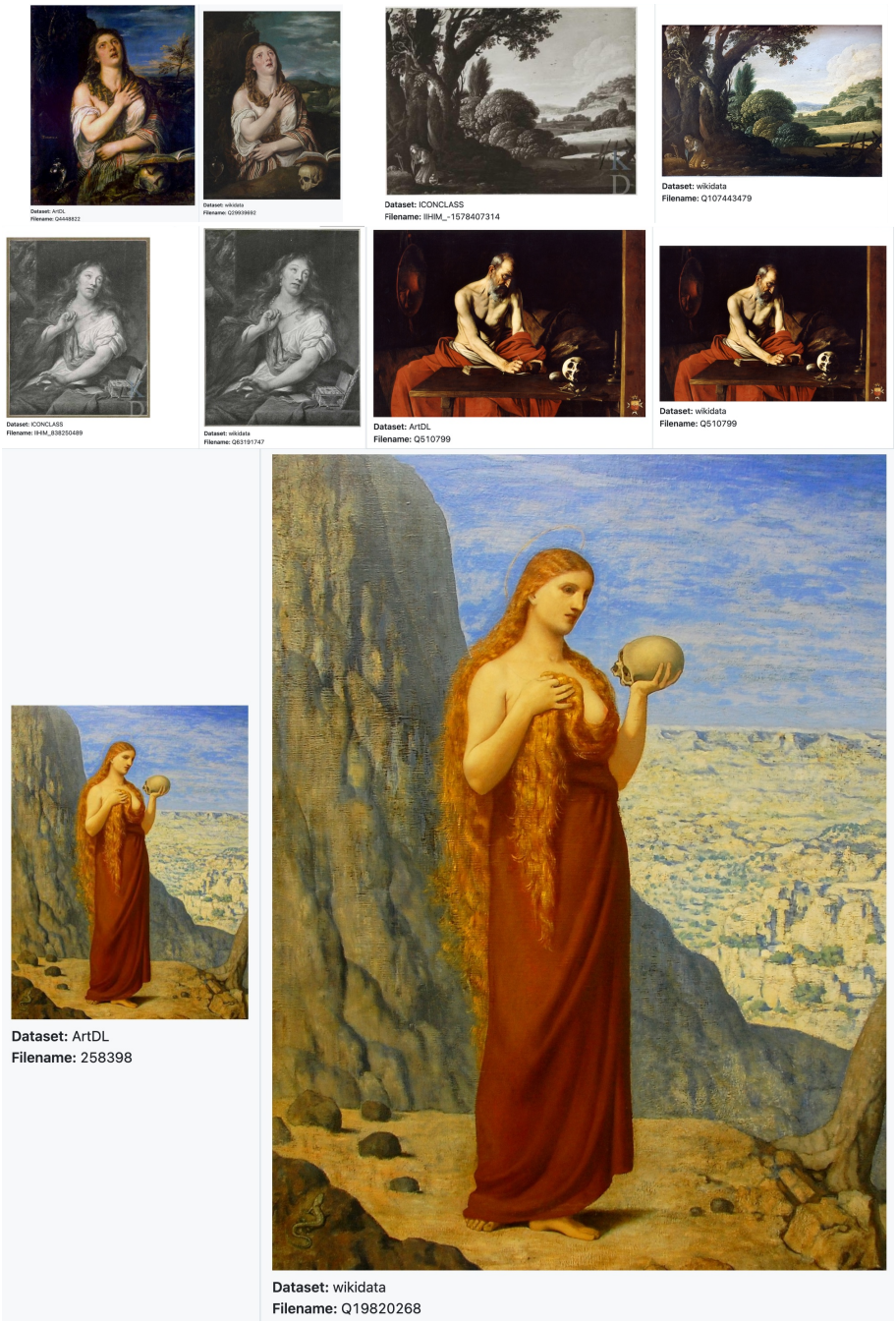

Figure 1: Example of image pairs deemed as similar (Hamming distance less than or equal to 8)

2.3.1 Technical implementation

The benchmarking pipeline consists of a set of Python scripts available on the GitHub repository[2], along with documentation that describes the detailed implementation points.

GPT-4o was accessed through the OpenAI APIs and configured to behave deterministically with a **temperature value of 0** and an arbitrary **seed of 12345**, and forced to produce a JSON output. The evaluation batches contained five images, which have been proven to be less error-prone and more cost-effective than other configurations. Prompts followed a fixed-based template and

[2] https://github.com/llm-art/ai-recognize-saints

dynamically included the list of class names or descriptions as candidate options[3]. These outputs were parsed using a structured JSON extraction routine, with fallbacks for occasional formatting anomalies.

Similarly, *Gemini 2.5* was tested using the Google AI Python SDK for the Gemini API. Likewise to GPT models, the **temperature was set to 0**, and the prompt was dynamically created by adding the list of candidate classes. These models have been configured with all safety categories (e.g., violent content, sexually explicit) set to *BLOCK NONE*, ensuring that religious artworks are not automatically filtered out or rejected if they contain commonly found religious themes, such as nudity or martyrdom.

These closed-source models are accessed via API endpoints and have transparency limitations in their training data, internal representations, and filtering mechanisms, of which we only know external hyperparameters such as tokens and context windows. To mitigate output variance and ensure consistency in classification, LLMs were configured to operate in a near-deterministic mode, exploiting specific hyperparameter values. While this reduces stochastic variation, it does not eliminate the nondeterminism of LLMs.

3 Results

This section presents the preliminary results of the benchmarking experiments, which are structured to assess the classification performance of the models for each test in a progressive manner. We begin by establishing a supervised baseline. To ensure the reliability of our evaluation, we then include a cross-dataset analysis, which serves as a consistency check and measures whether models consistently assign stable predictions to similar images across the datasets. We then report overall accuracy scores across all models, datasets, and test configurations for a fine-grained benchmarking on iconographic classification.

3.1 Baseline

To evaluate the models' performances, we established a baseline based on supervised CNNs using *ResNet50* architecture following the same methodology from the *ArtDL* paper [19]. We leveraged two models, which were explicitly fine-tuned on an 80% split of the images detected for the *ICONCLASS* and *Wikidata* datasets, respectively. The remaining 20% has been used for testing. On ICONCLASS, this approach achieved an accuracy of **40.46%**, while on Wikidata, it reached an accuracy of **43.97%**. Please refer to Appendix A for technical implementations and hyperparameters.

3.2 Cross-dataset consistency

We evaluate the robustness of model predictions by conducting a cross-dataset consistency analysis across **36 matched image pairs** in the three test scenarios. The goal was to measure the percentage of image pairs for which both images received identical model predictions. The results are shown in Table 1. The models demonstrated low consistency and greater sensitivity to the test configuration, particularly in terms of image sizes. *GPT-4o* ranged from **25.00%** in the first test to **27.78%** in the second and third tests. *GPT-4o-mini* had slightly lower overall scores but followed a similar trend. *Gemini-2.5-Flash* achieved consistency between **30.56%** and **33.33%**, while *Gemini-2.5-Pro* maintained a stable performance of **33.33%** across all tests, resulting in being the one with overall higher consistency.

[3] An example of prompt can be found on the GitHub repository linked above

Model	Test 1	Test 2	Test 3	Avg. Consistency
gpt-4o-2024-08-06	25.00%	27.78%	27.78%	26.85%
gpt-4o-mini-2024-07-18	22.22%	30.56%	25.00%	25.93%
gemini-2.5-flash-preview-05-20	30.56%	30.56%	**33.33%**	31.48%
gemini-2.5-pro-preview-05-06	**33.33%**	**33.33%**	33.33%	**33.33%**

Table 1: Consistency results across three tests for selected models, with the highest values in bold.

3.3 Classification performances

The classification performances are summarized in the following tables, showcasing the model's accuracy for the three tests: **(1)** is Zero-Shot with only labels, **(2)** is a Zero-Shot setting with Iconclass descriptions, and **(3)** is a Few-Shot approach with labels.

ArtDL Performances (Table 2) resulted similarly across models and configurations, ranging from **82.46%** for *GPT-4o-mini* in **Test 1** to a peak of **90.45%** achieved by *Gemini-2.5 Pro*. These models outperformed the baseline in almost all configurations. The baseline model achieved an accuracy of **84.44%**.

Model	Test 1	Test 2	Test 3
gpt-4o-2024-08-06	86.00%	87.45%	86.48%
gpt-4o-mini-2024-07-18	82.46%	84.98%	84.60%
gemini-2.5-flash-preview-05-20	88.20%	87.02%	84.71%
gemini-2.5-pro-preview-05-06	**90.45%**	**90.18%**	**86.59%**
Baseline	84.44%	84.44%	84.44%

Table 2: Accuracy scores across three tests for *ArtDL* dataset

For the ICONCLASS dataset (Table 3), *Gemini-2.5 Pro* achieved the highest performance, with accuracy scores of **83.31%**, **84.59%**, and **84.82%** across the three evaluation settings. *GPT-4o* showed stable performance across the settings. The baseline, fine-tuned on the *ICONCLASS* dataset, achieved an accuracy of **40.46%**.

Model	Test 1	Test 2	Test 3
gpt-4o-2024-08-06	75.32%	75.43%	73.46%
gpt-4o-mini-2024-07-18	55.74%	59.56%	55.50%
gemini-2.5-flash-preview-05-20	77.17%	77.75%	78.22%
gemini-2.5-pro-preview-05-06	**83.31%**	**84.82%**	**84.59%**
Baseline	40.46%	40.46%	40.46%

Table 3: Accuracy scores across three tests for *ICONCLASS* dataset

The *Wikidata* dataset (Table 4) generally showed lower accuracy scores compared to the other datasets, with most models clustering around the **35-45%** range. Among all models, *Gemini-2.5 Pro* achieved the highest performance, with a peak of **47.07%** in the third test. In contrast to their strong results on previous datasets, *GPT-4o* and *Gemini-2.5* models demonstrated more modest performance here. *GPT-4o-mini* fell behind the baseline, which reached **43.97%** accuracy, highlighting the difficulty of classifying this dataset for LLMs.

Across all three datasets, *Gemini-2.5 Pro* achieved the highest accuracy scores, consistently outperforming other models. On *Wikidata*, we also observed a general decline in accuracy. While

Model	Test 1	Test 2	Test 3
gpt-4o-2024-08-06	45.75%	45.31%	45.31%
gpt-4o-mini-2024-07-18	35.78%	36.95%	34.31%
gemini-2.5-flash-preview-05-20	45.45%	45.31%	44.57%
gemini-2.5-pro-preview-05-06	**45.89%**	**45.31%**	**47.07%**
Baseline	43.97%	43.97%	43.97%

Table 4: Accuracy scores across three tests for *ICONCLASS* dataset

top-performing models often exceeded **80%** accuracy on *ArtDL* and *ICONCLASS*, performance on *Wikidata* was relatively lower, reflecting a more challenging classification scenario. The *ResNet50* baselines were often outperformed by the models, despite being fine-tuned directly on the target datasets.

Finally, performance across the three evaluation configurations varied by dataset. On *ArtDL*, accuracy generally improved across the three stages, with few-shot learning leading to lower results, with only two cases, *Gemini 2.5 Flash* on *ICONCLASS* and *Gemini 2.5 Pro* on *wikidata* reaching **1-2%** increment in accuracy. On *ICONCLASS*, the pattern was less uniform but still showed improvement in several cases. On *Wikidata*, performance remained relatively stable across the three configurations, with no consistent trend of improvement.

4 Discussion

In this section, we discuss the performance of different model architectures and prompt designs in the task of classifying Christian iconography (**RQ1**) and how the results change after the progressive enrichment of contextual data (**RQ2**).

For **RQ1**, the results show that multimodal LLMs generally outperform traditional supervised models for the task of image classification of Christian iconography. For the three datasets, *Gemini-2.5 Pro* achieved the highest accuracy in all three evaluation settings, surpassing fine-tuned *ResNet50* baselines and yielding higher results in *ArtDL*. These findings confirm the effectiveness of multimodal LMMs in semantically dense and specific tasks. The Classification for *Wikidata* images witnessed overall lower accuracies, and this behavior suggests that these models tend to perform worse when facing inconsistent image qualities and sizes, such as the case of *Wikidata*, and is confirmed by the consistency check, where the identical image pairs are predicted differently when the sizes differ (Figure 2). Additionally, these LLMs are likely being trained on ArtDL and ICONCLASS datasets, as is common knowledge that the publishers scrape the Internet for training data. At the same time, Wikidata, which presents denser metadata, may not have its specific weights corresponding to the Iconclass codes tuned efficiently. Additionally, the two supervised *ResNet50* baselines, despite being fine-tuned only on a small set of 600 images, demonstrate that recent multimodal LLMs outperform supervised approaches, even for ArtDL, which has a more consistent number of images.

Regarding **RQ2**, we found that zero-shot classification using label descriptions generally improved accuracy except for *Wikidata*, where the changes are mostly negatives. Overall, the specific case of few-shot classification produced worse results. While *Gemini 2.5 flash* in *ICONCLASS* and *Gemini 2.5 Pro* in *Wikidata* demonstrated slight improvements in accuracy, the other models performed worse. This outcome proves that few-shot learning not always improve results and highlights the possible sensitivity to prompt formatting, class imbalance, or overfitting to a few visual characteristics. This also highlights the sensitivity to poorly chosen exemplars, which can degrade performance rather than enhance it.

Pair	Image 1	Dataset	Ground Truth	Predicted	Image 2	Dataset	Ground Truth	Predicted
1		ArtDL	11H(JOHN THE BAPTIST)	11H(JOHN THE BAPTIST)		wikidata	11H(JOHN THE BAPTIST)	11H(PAUL)
2		ArtDL	11H(PETER)	11H(PETER)		wikidata	11H(PETER)	11H(JOHN THE BAPTIST)
3		ArtDL	11H(PETER)	11H(PETER)		wikidata	11H(PETER)	11H(JOHN THE BAPTIST)
4		ArtDL	11H(JEROME)	11H(JEROME)		wikidata	11H(JEROME)	11HH(CATHERINE)

Figure 2: Example of incorrectly predicted images for Gemini 2.5 Pro, test 2. The image classification per row, one from ArtDL on the left and one from Wikidata. The focus is on the predicted class, differing for each dataset.

Despite these results, several limitations should be acknowledged. This study focuses on classifying single-labeled images, and the number of classes per dataset was relatively low; these limitations do not reflect a real-world scenario. Additionally, few-shot prompting did not consistently improve results, suggesting that the task is non-trivial. Further research is needed to optimize prompt formatting and choose the correct few-shot exemplars. These findings provide empirical support for using general-purpose multimodal models for the iconographical classification of artwork images. In particular, they validate the applicability of multimodal LLMs for low-data and high-complexity domains, such as Christian iconography, where semantic and label overlap is common. Even with these limitations, state-of-the-art models can detect and classify *Saint George* or *Saint Sebastian*, suggesting that these tools may help scholars in metadata curation without the need for extensive retraining.

5 Conclusions

This study serves to benchmark multimodal LLMs for the task of classifying Christian iconography, a domain with limited training data. The output is a reproducible evaluation framework spanning three datasets and classification settings, providing a baseline for future research in applying general-purpose AI systems to cultural heritage tasks. The results demonstrate that *Gemini 2.5 Pro* and *GPT 4o* can outperform traditional supervised baselines, indicating their potential as tools for metadata enrichment and semantic indexing in Digital Humanities workflows **(RQ1)**. Prompt enrichment improved performance in most settings, but few-shot learning mostly led to lower outcomes **(RQ2)**, suggesting that a more optimal example selection is required. This could

be addressed by integrating Retrieval Augmented Generation (RAG) pipelines while also reducing the risk of hallucinations. For future works, we aim to extend this benchmarking to include Vision-Language models, such as *CLIP*, *SigLIP*, and *BLIP-2*, which are commonly used for this task due to their high value in zero-shot classification. Additionally, to align this study with real-world scenarios, where classification is required for polyptychs or paintings featuring multiple saints, we need to integrate various visual and textual elements through training on datasets specifically providing this information.

Acknowledgements

The authors acknowledge Villa I Tatti, the Harvard University Center for Italian Renaissance Studies, for providing access to a dedicated computation server equipped with an NVIDIA L40S GPU (AD102GL architecture), which was essential for running vision-language model inference and fine-tuning of the baseline ResNet50 models. Villa I Tatti also provided API access to the GPT and Gemini 2.5 models, enabling the experiments presented in this work.

References

[1] Anthropic. “The Claude 3 Model Family: Opus, Sonnet, Haiku”. Claude-3 Model Card. 2024.

[2] Carriero, Valentina Anita, Gangemi, Aldo, Mancinelli, Maria Letizia, Marinucci, Ludovica, Nuzzolese, Andrea Giovanni, Presutti, Valentina, and Veninata, Chiara. “ArCo: The Italian Cultural Heritage Knowledge Graph”. In: *International Semantic Web Conference*. Auckland, New Zealand: Springer, 2019, pp. 36–52. DOI: https://doi.org/10.1007/978-3-030-30796-7_3.

[3] Castellano, Giovanna and Vessio, Gennaro. “Deep Learning Approaches to Pattern Extraction and Recognition in Paintings and Drawings: An Overview”. In: *Neural Computing and Applications 33*, no. 19 (2021), pp. 12263–12282. DOI: https://doi.org/10.1007/s00521-021-05893-z.

[4] Cetinic, Eva and She, James. “Understanding and Creating Art with AI: Review and Outlook”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 18*, no. 2 (2022), pp. 1–22. DOI: https://doi.org/10.1145/3475799.

[5] Cosovic, Marijana and Jankovic, Radmila. “CNN Classification of the Cultural Heritage Images”. In: *2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH)*. East Sarajevo, Bosnia and Herzegovina: IEEE, 2020, pp. 1–6. DOI: https://doi.org/10.1109/INFOTEH48170.2020.9066300.

[6] Couprie, L.D. “Iconclass: An Iconographic Classification System”. In: *Art Libraries Journal 8*, no. 2 (1983), pp. 32–49. DOI: https://doi.org/10.1017/S0307472200003436.

[7] Dosovitskiy, Alexey et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. arXiv preprint arXiv:2010.11929. 2020. DOI: https://doi.org/10.48550/arXiv.2010.11929.

[8] Gonthier, Nicolas, Gousseau, Yann, Ladjal, Said, and Bonfait, Olivier. “Weakly Supervised Object Detection in Artworks”. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. Munich, Germany, 2018. DOI: https://doi.org/10.1007/978-3-030-11012-3_53.

[9] Hawkins, Ashleigh. “Archives, Linked Data and the Digital Humanities: Increasing Access to Digitised and Born-digital Archives via the Semantic Web”. In: *Archival Science 22*, no. 3 (2022), pp. 319–344. DOI: https://doi.org/10.1007/s10502-021-09381-0.

[10] Inoue, Naoto, Furuta, Ryosuke, Yamasaki, Toshihiko, and Aizawa, Kiyoharu. “Cross-domain Weakly-supervised Object Detection through Progressive Domain Adaptation”. In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, USA, 2018, pp. 5001–5009. DOI: https://doi.org/10.1109/CVPR.2018.00525.

[11] Janković, Radmila. “Machine Learning Models for Cultural Heritage Image Classification: Comparison Based on Attribute Selection”. In: *Information 11*, no. 1 (2019), p. 12. DOI: https://doi.org/10.3390/info11010012.

[12] Jiang, Albert Q et al. “Mistral 7B”. arXiv preprint arXiv:2310.06825. 2023. DOI: https://doi.org/10.48550/arXiv.2310.06825.

[13] Kim, Jun-Hwa, Kim, Nam-Ho, Jo, Donghyeok, and Won, Chee Sun. “Multimodal Food Image Classification with Large Language Models”. In: *Electronics 13*, no. 22 (2024), p. 4552. DOI: https://doi.org/10.3390/electronics13224552.

[14] Labusch, Kai and Neudecker, Clemens. “Gauging the Limitations of Natural Language Supervised Text-Image Metrics Learning by Iconclass Visual Concepts”. In: *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*. 2023, pp. 19–24. DOI: https://doi.org/10.1145/3604951.3605516.

[15] Li, Junnan, Li, Dongxu, Savarese, Silvio, and Hoi, Steven. “BLIP-2: Bootstrapping Language-image Pre-training with Frozen Image Encoders and Large Language Models”. In: *International Conference on Machine Learning*. Honolulu, HI, USA: PMLR, 2023, pp. 19730–19742. DOI: https://dl.acm.org/doi/10.5555/3618408.3619222.

[16] Liu, Haotian, Li, Chunyuan, Wu, Qingyang, and Lee, Yong Jae. “Visual Instruction Tuning”. In: *Advances in Neural Information Processing Systems 36* (2023), pp. 34892–34916.

[17] Liu, Xiaolong, Deng, Zhidong, and Yang, Yuhan. “Recent Progress in Semantic Image Segmentation”. In: *Artificial Intelligence Review 52*, no. 2 (2019), pp. 1089–1106. DOI: https://doi.org/10.1007/s10462-018-9641-3.

[18] Maksimova, Erika, Meimer, Mari-Anna, Piirsalu, Mari, and Järv, Priit. “Viability of Zero-Shot Classification and Search of Historical Photos”. In: *Proceedings of the Computational Humanities Research Conference 2024*. Aarhus, Denmark, 2024, pp. 1242–1258.

[19] Milani, Federico and Fraternali, Piero. “A Dataset and a Convolutional Model for Iconography Classification in Paintings”. In: *Journal on Computing and Cultural Heritage 14*, no. 4 (2021), pp. 1–18. DOI: https://doi.org/10.1145/3458885.

[20] Mishra, Mayank and Lourenço, Paulo B. “Artificial Intelligence-assisted Visual Inspection for Cultural Heritage: State-of-the-art Review”. In: *Journal of Cultural Heritage 66* (2024), pp. 536–550. DOI: https://doi.org/10.1016/j.culher.2024.01.005.

[21] O’shea, Keiron and Nash, Ryan. “An Introduction to Convolutional Neural Networks”. arXiv preprint arXiv:1511.08458. 2015. DOI: https://doi.org/10.48550/arXiv.1511.08458.

[22] Panofsky, Erwin. *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. London: Routledge, 2018.

[23] Posthumus, Etienne. “Iconclass AI Test Set”. Online dataset. 2020.

[24] Radford, Alec, Narasimhan, Karthik, Salimans, Tim, and Sutskever, Ilya. “Improving Language Understanding by Generative Pre-training”. OpenAI Technical Report. 2018.

[25] Radford, Alec et al. “Learning Transferable Visual Models from Natural Language Supervision”. In: *International Conference on Machine Learning*. Virtual: PMLR, 2021, pp. 8748–8763.

[26] Santini, Cristian, Posthumus, Etienne, Tietz, Tabea, Tan, Mary Ann, Bruns, Oleksandra, and Sack, Harald. “Multimodal Search on Iconclass using Vision-Language Pre-Trained Models”. In: *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE. 2023, pp. 285–287. DOI: https://doi.org/10.1109/JCDL57899.2023.00061.

[27] Sartini, Bruno. “IICONGRAPH: Improved Iconographic and Iconological Statements in Knowledge Graphs”. In: *European Semantic Web Conference*. Hersonissos, Greece: Springer, 2024, pp. 57–74. DOI: https://doi.org/10.1007/978-3-031-60635-9_4.

[28] Sartini, Bruno, Baroncini, Sofia, Erp, Marieke van, Tomasi, Francesca, and Gangemi, Aldo. “ICON: An Ontology for Comprehensive Artistic Interpretations”. In: *ACM Journal on Computing and Cultural Heritage 16*, no. 3 (2023), pp. 1–38. DOI: https://doi.org/10.1145/3594724.

[29] Springstein, Matthias, Schneider, Stefanie, Rahnama, Javad, Stalter, Julian, Kristen, Maximilian, Müller-Budack, Eric, and Ewerth, Ralph. “Visual Narratives: Large-Scale Hierarchical Classification of Art-Historical Images”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2024, pp. 7220–7230. DOI: https://doi.org/10.1109/WACV57701.2024.00705.

[30] Steinebach, Martin. “Robust Hashing for Efficient Forensic Analysis of Image Sets”. In: *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 180–187. DOI: https://doi.org/10.1007/978-3-642-35515-8_15.

[31] Team, Gemini et al. “Gemini: A Family of Highly Capable Multimodal Models”. arXiv preprint arXiv:2312.11805. 2025. URL: https://arxiv.org/abs/2312.11805.

[32] Yu, Qing and Shi, Ce. “An Image Classification Approach for Painting Using Improved Convolutional Neural Algorithm”. In: *Soft Computing 28*, no. 1 (2024), pp. 847–873. DOI: https://doi.org/10.1007/s00500-023-09420-1.

[33] Zhai, Xiaohua, Mustafa, Basil, Kolesnikov, Alexander, and Beyer, Lucas. “Sigmoid Loss for Language Image Pre-training”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Paris, France, 2023, pp. 11975–11986.

A Training Parameters for Baseline Models

The supervised baseline models were trained using *ResNet50* architecture with the following hyperparameters for both datasets.

The models were trained using weighted sampling to address class imbalance in both datasets. Training was performed using PyTorch with *Adam optimizer* and *cross-entropy loss*. Early stopping was applied based on validation accuracy with a patience of 10 epochs.

Parameter	Hyperparameters
Epochs	50
Batch Size	32
Learning Rate	1e-3
Image Size	224×224
Pretrained Weights	ImageNet
Weighted Sampling	Yes

Table 5: Training hyperparameters for baseline *ResNet50* models on ICONCLASS and Wikidata datasets

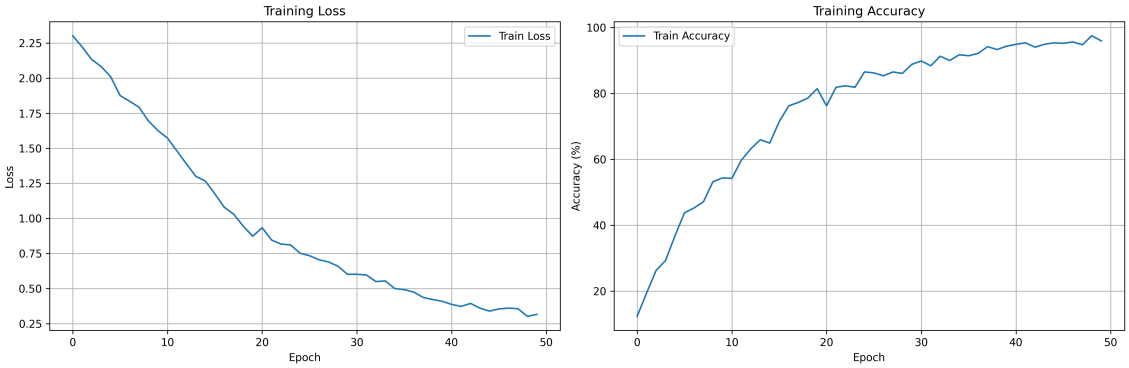

Figure 3: Training and validation curves for the *ResNet50* baseline model on the ICONCLASS dataset. The plots show accuracy and loss progression over 50 epochs with the hyperparameters specified in Table 5.

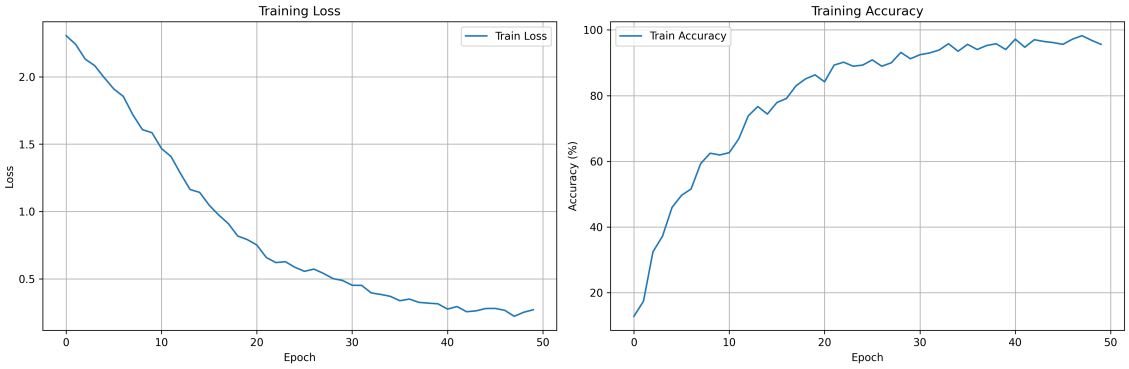

Figure 4: Training and validation curves for the *ResNet50* baseline model on the Wikidata dataset. The plots show accuracy and loss progression over 50 epochs with the hyperparameters specified in Table 5.