histories

MDPI

Article

Transferring AI-Based Iconclass Classification Across Image Traditions: A RAG Pipeline for the *Wenzelsbibel*

Drew B. Thomas [1,*] and Julia Hintersteiner [2]

[1] Department of History, University of Salzburg, 5020 Salzburg, Austria
[2] Department of German Studies, University of Salzburg, 5020 Salzburg, Austria; julia.hintersteiner@plus.ac.at
* Correspondence: drew.thomas@plus.ac.at

Abstract

This study evaluates whether a multimodal retrieval-augmented generation (RAG) pipeline originally developed for early modern woodcuts can be effectively transferred to the domain of medieval manuscript illumination. Using a dataset of *Wenzelsbibel* miniatures annotated with Iconclass, the pipeline combined page-level image input, LLM description generation, vector retrieval, and hierarchical reasoning. Although overall scores were lower than in the earlier woodcut study, the best-performing configuration still substantially surpassed both image-similarity and keyword-based search, confirming the advantages of structured multimodal retrieval for medieval material. Truncation analysis further revealed that many errors occurred only at the deepest Iconclass levels: removing levels raised precision to 0.64 and 0.73, with average remaining depths of 5.49 and 4.49 levels, respectively. These results indicate that the model's broader hierarchical placement is often correct even when fine-grained specificity breaks down. Taken together, the findings demonstrate that a woodcut-oriented RAG pipeline can be meaningfully adapted to manuscript illumination and that its strengths lie in contextual reasoning and structured classification. Future improvements should incorporate available textual metadata, explore graph-based retrieval, and refine Iconclass-driven pathways.

Keywords: Iconclass classification; medieval manuscripts; manuscript illumination; biblical illustration; retrieval-augmented generation; multimodal models; digital art history; scene recognition; knowledge organization systems; digital humanities

Academic Editors: Matteo Valleriani and Doris Gruber

Received: 1 December 2025
Revised: 5 February 2026
Accepted: 11 February 2026
Published: 18 February 2026

1. Introduction

1.1. Background and Research Objectives

Digitized cultural heritage collections often contain thousands of images that remain difficult to search, compare, or analyze at scale because their iconographic content has never been consistently described. Assigning structured vocabularies such as Iconclass (Brandhorst and Posthumus) is one of the most effective ways to make historical images interoperable, but producing these classifications manually requires considerable time, training, and domain-specific judgment (Posthumus et al. 2022; Santini et al. 2024). This is especially true for richly illustrated biblical manuscripts, where visual narratives can be layered or ambiguous, and where even small stylistic differences may carry interpretive significance.

The *Wenzelsbibel*, a lavishly illuminated late 14th-century German Bible, offers a clear example of this challenge. It has hundreds of miniatures depicting biblical narratives, but most have never been indexed with controlled iconographic vocabulary. As part

of an ongoing digital humanities project at the Paris Lodron University of Salzburg, a portion of these scenes has been classified using Iconclass, providing a ground-truth dataset (detailed below). This makes it possible to test whether recent developments in AI-assisted classification could help accelerate or support such work without replacing the expertise upon which it depends.

The present study builds on an AI pipeline originally developed for early modern biblical woodcuts. That earlier work, part of the "Visualizing Faith" project at University College Dublin (Thomas 2022–2026), demonstrated that combining multi-modal Large Language Models (LLMs) and vector search with Retrieval-Augmented Generation (RAG) enabled automatic classification of Iconclass codes with promising levels of consistency (Thomas 2026). Transferring the methodology to the *Wenzelsbibel* offered an opportunity to evaluate not only the technical performance of the pipeline but also the adaptability of the approach across different artistic media, periods, and interpretive contexts.

This article reports on that collaborative experiment. On the one hand, it describes the construction and refinement of a multi-phase AI workflow: vector retrieval of candidate Iconclass codes, hierarchical expansion of parent concepts, a calibrated "none" option, and a rescue pipeline designed to correct for recurrent misclassification patterns. On the other hand, it draws on expert-supplied ground truth to evaluate the system's behavior, to identify where it aligns with scholarly judgment, and to understand where it fails or requires human correction. Rather than testing whether AI can replace manual classification, we aim to assess whether such a pipeline can support human expertise by speeding up initial annotation, suggesting plausible alternatives, and increasing the consistency of large-scale metadata work.

By focusing on the concrete task of classifying the *Wenzelsbibel's* miniatures, this study contributes to a growing body of research evaluating how AI can be integrated into historical and art-historical workflows. More broadly, it demonstrates the importance of collaborative experimentation between historians, digital scholars, and AI systems.

1.2. Related Work

This study builds directly on previous work that introduced the underlying methodology and evaluated it on early modern woodcuts (Thomas 2026); the present paper, therefore, does not repeat that discussion in full, but briefly situates the approach within the broader landscape of AI-based cultural heritage research and recent work on Iconclass automation.

Recent research has demonstrated the growing applicability of computer vision and vision–language models to the analysis of cultural heritage imagery, particularly for improving access to large-scale visual collections. In the domain of early modern books and prints, object detection and image analysis techniques have been used to identify and organize illustrations, woodcuts, and printers' ornaments across extensive corpora, enabling new forms of large-scale visual comparison and discovery (Bergel et al. 2013; Wilkinson et al. 2021; Wilkinson 2022)

At the same time, numerous studies have emphasized the challenges of transferring models trained on modern images to historical artworks, which frequently depict symbolic, allegorical, or non-naturalistic subjects and lack the large, richly annotated datasets required for supervised learning. In response, several initiatives have explored the combination of visual features with structured metadata or external knowledge sources, as well as the creation of large benchmark datasets and exploration platforms for cultural heritage imagery (Reshetnikov et al. 2022; Strezoski and Worring 2018; Springstein et al. 2021).

Within this broader landscape, increasing attention has been given to the automation of iconographic classification using controlled vocabularies such as Iconclass. The Iconclass project itself has recently introduced semantic and image-based search functionality into its

online interface with the Iconclass Plus browser (Posthumus et al. 2022), and recent research has investigated multimodal retrieval for predicting Iconclass codes, often highlighting the benefits of combining textual and visual representations alongside the computational costs of fine-tuning neural networks (Santini et al. 2024; Banar et al. 2021, 2023).

2. Materials and Methods

2.1. Corpus

The *Wenzelsbibel* is one of the most ambitious and visually sophisticated German biblical manuscripts of the late 14th century (Österreichische Nationalbibliothek n.d., Cod. 2759–2764). Commissioned for King Wenceslas IV of Bohemia and produced around 1390–1400, it comprises six volumes, 1214 folios, and a richly conceived program of biblical illustration, which remains unfinished (Hlaváčková 2001). Beyond its artistic value, the *Wenzelsbibel* is historically significant as the earliest complete German translation of the Old Testament based on the Latin Vulgate, presented alongside an extensive and visually coordinated cycle of illuminations (see Figure 1). Its scale and narrative density make it a remarkable achievement of late medieval biblical culture (Theisen 2024).

Figure 1. The End of the Flood: Noah in the Ark as the dove brings him a branch with green leaves. Vienna, ÖNB, Cod. 2759, f. 8v.

The manuscript is the focus of the ongoing digital edition *Die Wenzelsbibel—Digitale Edition und Analyse*, led by Prof. Manfred Kern at the University of Salzburg in cooperation with the Austrian National Library (Kern 2022–2024). The edition unites high-resolution facsimiles, diplomatic and normalized transcriptions, text-critical notes, commentary, and image analysis within a unified TEI/XML and RDF framework. Each folio is treated as an integrated visual-textual ensemble, capturing not only the biblical text but also marginal inscriptions, painting instructions, and the full hierarchy of miniature, initial, and ornamental decoration. In the long term, the edition will include all six volumes; at present, the Pentateuch (Genesis–Deuteronomy) is the portion for which the image analysis and commentary have been completed.

The Pentateuch corpus contains 230 miniatures across 226 folios. Many of these miniatures depict more than one biblical episode, either through pluriscenic composition, narrative compression, or multi-register layouts. Specifically, several miniatures are divided into two, with top and bottom scenes, respectively. For evaluation, such miniatures were subdivided into analytically distinct units. This resulted in a working dataset of 277 scenes, each aligned

with a specific page, contextualized by its surrounding biblical text, and linked to one or more manually assigned Iconclass codes (see Figure 2; Thomas and Hintersteiner 2026).

Figure 2. Manuscript page showing The End of the Flood: Noah in the Ark as the dove brings him a branch with green leaves. Vienna, ÖNB, Cod. 2759, f. 8v.

From the perspective of AI-assisted classification, the Pentateuch miniatures present a rich but challenging corpus. The imagery ranges from straightforward monoscenic depictions to complex composite compositions, as figures may appear repeatedly within a single frame. Additional challenges arise when Deuteronomy recounts earlier events from Exodus, Leviticus, or Numbers, resulting in illustrations whose Iconclass categories belong to different parts of the hierarchy. These conditions make the dataset an excellent test case for assessing whether an AI pipeline developed for early modern printed woodcuts can be successfully transferred to the materially and visually distinct domain of medieval illumination.

2.2. Iconclass

Iconclass, developed by Henri van de Waal in the 1940s, is a widely used system for classifying historical images through a controlled, alphanumeric vocabulary (van de Waal 1974–1985). Its strength lies in its hierarchical structure: specific visual motifs or narrative scenes inherit from broader conceptual categories. Each additional character in an Iconclass code narrows the scope, leading from general categories (e.g., "Bible") to highly specific narrative moments.

A typical example illustrates this structure clearly:

7. Bible
71. Old Testament
71B. Genesis from the descendants of Cain and Seth to Abraham
71B3. The story of Noah
71B33. The Flood and destruction of mankind ~ story of Noah (Genesis 7:10–8:17)
71B332. The Ark of Noah comes to a rest on Mount Ararat
71B3322. Noah sends off a dove
71B33221. The dove returns with an olive-branch ~ story of Noah

This hierarchy demonstrates how Iconclass encodes narrative meaning at multiple levels. A scholar might be interested only in a general category (e.g., the story of Noah), in a specific subsection (e.g., all scenes after the Flood's destruction), or in the narrowly defined moment where the dove returns with the olive branch. Iconclass enables retrieval across all these levels because each inherits from the categories above it.

For large manuscripts, this design is enormously helpful. A project may want to annotate a miniature with a precise code, such as 71B33221, but a researcher interested only in the story of Noah could retrieve the miniature simply by searching at the level of 71B3. Conversely, broad categories can be used when the miniature lacks sufficient detail to justify a specific classification.

However, several characteristics of medieval biblical illustration expose structural limitations in the Iconclass hierarchy:

1. *Strict hierarchical alignment can conflict with biblical intertextuality*. The biblical narrative frequently refers to earlier events, especially in books like Deuteronomy, where Moses recounts events that took place in preceding books. As a result, a miniature placed in Deuteronomy may depict an event whose canonical narrative origin lies in Exodus, Leviticus, or Numbers. While such an image may correspond exactly to a specific Iconclass code under the hierarchy of the earlier book, assigning it there would misrepresent its manuscript context. For example, the defeat of King Sihon by the Israelites is illustrated in Deuteronomy on fol. 177r, where Moses retells the episode, even though the event itself occurs in the book of Numbers and is classified accordingly in Iconclass (71E326(+0)). Iconclass mixes narrative categories and textual-structural categories, and these do not always match the ways medieval illustrators organized their illustrations.

2. *Not all miniatures have a precise Iconclass equivalent.* While many scenes map directly to established narrative codes, others are not explicitly covered in the Iconclass vocabulary. In these cases, only a higher-level or more generic category can be applied.

3. *Fragmentation between "Religion and Magic" (1) and "Bible" (7) categories*. Biblical material is not confined to the 7 ("Bible") subtree. Many relevant concepts appear under 1 ("Religion and Magic"), especially ritual motifs such as the Tabernacle, the Ark of the Covenant, or liturgical furnishings. A miniature might need a broad 7-category signaling its book-level context (e.g., "Exodus"), and a 1-category identifier for a specific object or symbolic motif. This fragmentation complicates retrieval and creates ambiguity about which branch to prioritize.

Together, these constraints make the *Wenzelsbibel* a revealing test case for automated classification. Its miniatures fully engage the strengths of Iconclass (narrative hierarchy), while exposing areas where its structure requires supplementing with expert interpretation. They also provide exactly the kinds of ambiguous, multi-layered inputs that highlight the limits—and potential—of large language models and retrieval pipelines in art-historical classification work.

The methodology used in this study builds directly on a pipeline originally developed to classify early modern biblical woodcuts. That earlier project explored how image descriptions generated by LLMs could be combined with vector search to identify the most likely Iconclass codes for a given illustration. It tested numerous configurations, including whether the model performed better when shown the cropped illustration alone or the entire printed page. Because the printed pages usually contain headings, chapter numbers, running text, and other contextual cues, the study found that the model produced more accurate and contextually grounded descriptions when given the full-page image. This in turn improved the retrieval stage of the pipeline. The earlier study also compared traditional keyword search, vector search, and hybrid search, and evaluated two Weaviate vector databases: one containing the Iconclass descriptions only, and one enriched with all

parent-level Iconclass codes. It further incorporated an image-search baseline using the Iconclass AI Test Set (Posthumus 2020), which allowed for comparison between vision-only retrieval and the multimodal RAG pipeline.

From these extensive experiments, two configurations consistently outperformed the others: a vector-search pipeline using the basic Iconclass database and a hybrid-search pipeline using the hierarchical database. Importantly, both top-performing approaches relied on full-page images for description generation. These two configurations therefore formed the basis for the present study. The central question was whether an approach optimized for early modern prints could be transferred to the domain of medieval manuscript illumination.

2.3. Phase 1: Testing Transferability and Model Variants

Phase 1 of the *Wenzelsbibel* study focused on testing the most important variables identified in the earlier work, but now applied to medieval miniatures rather than printed woodcuts. The phase compared image-only input to full-page input in order to determine whether an LLM could extract meaningful information from medieval scripts, layout, and rubrication. Whereas full-page context had significantly improved description quality in early modern material, it could not be assumed that the model would be equally capable of interpreting medieval handwriting or manuscript paratexts.

Phase 1 also compared two LLM models, GPT-4o and GPT-4.1, for their ability to generate accurate and detailed descriptions of the images, and employed two prompt variants, one stating that the input image was a "Biblical illustration" and another that it was specifically an "Old Testament illustration." The retrieval component (vector or hybrid search) used updated embeddings (text-embedding-3-small, replacing the earlier OpenAI text-embedding-ada-002 embedding model). All Phase 1 configurations ended with a simple RAG step: the model was presented with the top five Iconclass results returned by the vector database and asked to choose the best match. This created a consistent and controlled experimental environment for determining which combination of inputs, models, prompts, and retrieval strategies transferred most effectively from early modern prints to medieval illuminations.

Phase 1 produced a clear best-performing configuration, and this was carried forward into Phase 2.

2.4. Phase 2.1: Incorporating Iconclass Hierarchies into the RAG Step

Phase 2 developed the pipeline further by introducing hierarchical reasoning. Iconclass is inherently hierarchical, and the earlier study had confirmed that parent-level information can help an LLM discriminate between narrowly specific codes and more general ones. In Phase 2.1, the five vector results from Weaviate were expanded to include their parent codes. Two representations were tested. In the first, the parent codes were presented as a flat list ordered simply by Iconclass number. In the second, the same parent codes were arranged as a hierarchical tree using indentation and visual markers to indicate the structure of the Iconclass system. These parent-expanded representations were then supplied to the LLM, which was asked to choose the most appropriate Iconclass code from the combined list.

However, an unexpected issue emerged: sorting the expanded parent lists by Iconclass code inadvertently erased the original ordering of the top-five vector results. The ranking of these results is meaningful—the RAG step exists, in part, to correct cases where the top-ranked vector result is not actually the best match. But once parent expansion and sorting were applied, the model no longer had access to the original retrieval confidence order. To solve this problem, Phase 2.1b reintroduced the ranking context by providing the LLM with both the hierarchical parent tree and a separate section listing the original top-five vector

hits in ranked order. This ensured that the model had access to both semantic structure and relevance weighting.

The best-performing configuration from Phase 2.1 was then selected for further refinement in Phase 2.2.

2.5. Phase 2.2a: Introducing a "None" Option and the Rescue Pipeline

A recurring pattern in Phases 1 and 2.1 was that the model sometimes selected an Iconclass code even when none of the candidates were appropriate. To mitigate this, Phase 2.2 introduced an explicit "None of the above" option. The model was instructed that if none of the presented Iconclass codes were correct, it should respond with "None." This change enabled a multi-step rescue pipeline designed to recover from poor initial retrievals.

The rescue pipeline in its first version (Phase 2.2a) began by examining cases where the model had selected "None." One discovery was that occasional scenes that clearly belonged within the biblical 7 Iconclass category nonetheless retrieved only 1 (Religion and Magic) codes from the vector search. In these cases, even though the vector search produced five results, the set was systematically biased. Step 2 of the rescue pipeline, therefore, checked the prefixes of all five candidates: if all were from category 1, the pipeline executed a new vector search restricted to 7, and vice versa. This ensured that the model was always given the opportunity to consider results from the correct portion of the Iconclass hierarchy. If the new filtered search produced candidates, a second RAG decision followed.

If the scene still could not be confidently classified, the pipeline proceeded to a broader selection stage. At this point, the model was no longer asked to choose a narrowly specific Iconclass code but was instead directed to choose a higher-level parent code that represented the general narrative category of the scene. This recognized the principle that a broad but correct identification is preferable to none at all.

If even this broader classification failed, the pipeline moved to a book-level rescue stage. Here the model attempted to identify, at minimum, the biblical book illustrated—Genesis, Exodus, Leviticus, Numbers, or Deuteronomy. This step acknowledged that although the model might be unable to identify the specific scene, it could still often detect the position within the biblical narrative. If this too proved impossible, the pipeline resorted to a controlled fallback: the code 71, representing the Old Testament in Iconclass. This ensured that every image received a classification, even if only at the most general level.

2.6. Phase 2.2b: A More Selective Rescue Strategy

Phase 2.2b refined the rescue pipeline in two important ways. First, it removed the broadest parent-level Iconclass codes from consideration during the early rescue stages. Because later steps in the pipeline already addressed broad biblical book and Old Testament identifiers, retaining such high-level parents earlier in the process diluted the specificity of the RAG decisions. Their removal ensured that the model's early rescue attempts focused on genuinely informative mid-level Iconclass categories.

Second, Phase 2.2b adjusted how the parent codes were grouped. Instead of presenting all available parent codes together, the pipeline separated them into those beginning with 7 (Bible) and those beginning with 1 (Religion and Magic). Based on the distribution of scenes in the *Wenzelsbibel*, the 7 hierarchy is far more likely to contain the correct classification. For this reason, the pipeline was modified to always try the 7-group parent codes first, only falling back to 1-group codes if the biblical subtree failed. This change reduced misclassification caused by structurally plausible but contextually inappropriate 1 codes.

Across these stages, the methodology evolved from a straightforward top-five RAG selection into a sophisticated, hierarchical, and context-aware classification system. The transfer tests from the earlier woodcut study provided the foundation; Phase 1 adapted

these results to the distinctive characteristics of medieval manuscript imagery; Phase 2 introduced hierarchical reasoning and new logic for recovery when retrieval failed. Collectively, these refinements allowed the pipeline not only to identify specific Iconclass codes but also to handle uncertainty gracefully, resolve inconsistent retrieval patterns, and provide appropriate fallback solutions rooted in the structure of the biblical narrative.

3. Evaluation

The evaluation framework was designed to assess both the accuracy and hierarchical appropriateness of the model's Iconclass predictions. Because Iconclass classifications are inherently layered and can vary in specificity, the evaluation combined exact-match metrics, a weighted hierarchical scoring system, and a truncation-based analysis. Together, these methods allow us to judge not only whether the model arrived at the "right" code, but also how close it came when it missed, and whether it nonetheless located the correct conceptual region within the Iconclass structure.

A crucial preliminary step in the evaluation was defining the ground truth against which predictions would be compared. Each scene in the *Wenzelsbibel* dataset was assigned one or more primary ground truth Iconclass codes. These represent the correct classifications as determined through expert analysis of the image's narrative content. In several cases, multiple codes may be appropriate. These codes constitute the authoritative reference labels for evaluation.

In addition, a small subset of scenes (sixteen in total) was assigned secondary ground truth codes. These were used in cases where a miniature depicts a biblical episode that formally belongs to a different book of the Bible. For example, a scene in Leviticus may visually recount an event originally narrated in Exodus. Iconclass encodes these episodes within the Exodus branch of its hierarchy, meaning that the precise Iconclass for the depicted scene would be nested under the Exodus tree. Applying that code to an illumination in Leviticus would technically misrepresent its biblical location, yet the code remains narratively accurate when the scene is viewed in isolation. Because these secondary codes convey genuine iconographic correctness despite their structural mismatch within Iconclass, they were included as valid alternative matches. A prediction was counted as correct if it matched any of the primary or secondary ground truth labels associated with the scene.

With the ground truth defined, the first set of evaluation metrics consisted of precision, recall, and F1 score. For each scene, the model produced a single predicted Iconclass code. Precision measured the proportion of those predictions that matched at least one of the ground truth codes. Recall measured the proportion of scenes for which the model successfully retrieved one of the relevant codes. The F1 score combined both measures into a single harmonic mean. These metrics allow for direct comparison with the earlier study on early modern woodcuts and provide a baseline assessment of exact classification accuracy.

Exact matching, however, captures only part of the picture. Iconclass is a hierarchical system in which a prediction may be "wrong" at the deepest level yet still correctly identify the broader narrative category. To capture these intermediate degrees of correctness, the evaluation employed a weighted scoring system that assigns partial credit based on how closely the predicted code aligns with the ground truth hierarchy. When the model's prediction exactly replicated the ground truth code, it received the highest score. When the prediction recovered the full ground truth structure but added unnecessary or overly specific levels, it received a slightly lower score reflecting that the additional detail may be speculative even if the underlying classification was sound.

Other forms of partial matches were scored according to how many levels in the predicted code corresponded to the correct hierarchical path. Predictions that captured the broad narrative grouping but not the specific episode were recognized as largely correct,

while predictions that overlapped only partially with the ground truth hierarchy received proportionally lower scores. The scoring system incorporated a proportional adjustment that scaled the value of a prediction according to the share of hierarchical levels it correctly matched. Conversely, when the predicted code introduced incorrect branches—that is, levels that contradicted the ground truth—penalties were applied, though not to the extent that meaningful partial matches were rendered valueless. This weighted approach provides a fine-grained assessment of the model's capacity to approximate the correct classification even when an exact match is not achieved.

Finally, the evaluation included hierarchical truncation as a way of examining how well the model performed at broader conceptual levels. After generating predictions at full depth, the lowest level of each predicted Iconclass was removed, and the evaluation metrics were recalculated. This process was repeated iteratively. Truncation exposes how often the model succeeded in placing an image within the correct branch of the iconographic hierarchy tree, even when it failed to identify the exact leaf-level code. In the earlier woodcut study, truncation revealed that many apparent errors were, in fact, correct at broader hierarchical levels. Applying truncation to the *Wenzelsbibel* data allowed us to assess whether similar behavior emerged in the medieval context.

Together, the combination of exact-match metrics, weighted hierarchical scoring, and truncation-based evaluation provides a comprehensive understanding of how the model performs across the full range of Iconclass specificity. This multi-layered approach is particularly important for complex manuscripts such as the *Wenzelsbibel*, where miniatures may compress or juxtapose multiple scenes, and where a partially correct broader identification can remain highly useful for cataloging and research.

4. Results

4.1. Baseline Performance

4.1.1. Image-Based Baseline: OpenCLIP Nearest-Neighbor Search

The first baseline evaluated the performance of a purely vision-based method using OpenCLIP to perform nearest-neighbor retrieval against the Iconclass AI Test Set. For each *Wenzelsbibel* scene, the system retrieved the ten most visually similar images from the reference corpus and predicted the Iconclass code most frequently associated with these neighbors. This method follows the logic of the Iconclass+ interface and provides a strict vision-only comparator, devoid of textual or hierarchical reasoning.

The results confirm that visual similarity alone provides limited guidance for classifying medieval manuscript illumination. The OpenCLIP baseline achieved a precision of 0.1694, a recall of 0.1131, and an F1 score of 0.1310. These low scores indicate that the nearest-neighbor method frequently retrieved images that were visually analogous—similar composition, color, or pose—yet semantically incorrect.

The weighted hierarchical scoring system paints a similar picture. The OpenCLIP baseline achieved an average weighted score of 22.76, indicating that even when predictions did not match the ground truth exactly, they rarely aligned well with the correct branch of the Iconclass hierarchy. This score reflects that most predictions failed not only at the specific level but also at capturing the broader narrative grouping.

Together, these metrics confirm that a purely visual retrieval method cannot reliably classify *Wenzelsbibel* scenes. The baseline serves as a useful lower bound: any successful RAG-based pipeline must substantially outperform this vision-only approach, both in exact matching and in broader hierarchical alignment.

4.1.2. Text-Based Baseline: BM25 Keyword Search

The second baseline relied solely on traditional keyword matching using BM25 (Best Matching 25), a probabilistic ranking function widely used in information retrieval (Robertson et al. 1995). In this setup, the model-generated image descriptions served as queries against the Iconclass database, and entries were ranked by lexical similarity. Unlike the OpenCLIP baseline, this method incorporates no visual reasoning and no awareness of the Iconclass hierarchy beyond what is encoded in the textual descriptions. It therefore provides a useful comparator for how far textual matching alone can go when supplied with reasonably detailed image descriptions.

BM25 performed significantly better than the image-based baseline. The keyword search achieved a precision of 0.462, a recall of 0.547, and an F1 score of 0.494. These scores indicate that many descriptions contain sufficient lexical cues—names of figures, references to locations or objects, or scriptural terms—to retrieve relevant Iconclass entries even without semantic embeddings. The relatively strong recall suggests that BM25 often located the correct general region of the Iconclass vocabulary, although precision remained modest due to cases where key terms were either absent from the database entry or misleadingly matched to related but incorrect concepts.

The weighted hierarchical scoring system further reflects this pattern. BM25 obtained an average weighted score of 38.11, substantially higher than the OpenCLIP baseline. The weighted score suggests that BM25 often retrieved codes that were directionally correct in the hierarchy but lacked the specificity required for exact matches. These results reinforce the need for a combined approach that integrates vision, language, retrieval, and hierarchical reasoning, as implemented in the subsequent RAG-based phases.

4.2. Phase 1 Results: Transferability from Woodcuts to Illuminations

Phase 1 evaluated the extent to which the best-performing configurations from the earlier woodcut study transferred to the *Wenzelsbibel* miniatures. To illustrate the type of textual input generated by the LLM at this stage, the following is an example of a model-generated description of the creation of Eve on f. 4r:

> "The illustration depicts God creating Eve from Adam's side as Adam lies on the ground, with vibrant trees in the background suggesting the Garden of Eden. God, shown with a halo and wearing flowing robes, lifts Eve up from Adam's side as she emerges." (Full page, GPT-4.1, Old Testament prompt)

The variables tested included image-only versus full-page input, the use of GPT-4o versus GPT-4.1 for description generation, two prompt variants ("Biblical illustration" versus "Old Testament illustration"), and two retrieval configurations (basic vector search and hierarchical hybrid search).

Across all configurations, one pattern immediately mirrored the findings of the woodcut study: full-page input consistently outperformed image-only input (see Table 1). This suggests that even in medieval manuscripts, where scripts may be visually more challenging than early modern type, the model was nevertheless able to extract and utilize contextual cues such as headings or layout structure. Although the benefit was more modest than in printed books, the difference was systematic across models, prompts, and retrieval methods.

Another clear pattern was the performance gap between the two LLMs. In every configuration, GPT-4.1 outperformed GPT-4o, both in precision and recall. This held true for both image-only and full-page input, indicating that the improved multimodal capabilities of GPT-4.1 translated meaningfully to the medieval manuscript domain.

Table 1. Results from Phase 1.

Prompt	View	LLM	Vector/RAG Model	Score	Precision	Recall	F1
Bible	Image	GPT-4.1	Basic Vector	41.08	0.4439	0.5036	0.4651
Bible	Image	GPT-4.1	Hierarchical Hybrid	39.46	0.4167	0.4919	0.4444
Bible	Image	GPT-4o	Basic Vector	35.91	0.3764	0.4286	0.3943
Bible	Image	GPT-4o	Hierarchical Hybrid	35.75	0.3604	0.4133	0.3793
Bible	Page	GPT-4.1	Basic Vector	43.02	0.4993	0.5760	0.5265
Bible	Page	GPT-4.1	Hierarchical Hybrid	42.18	0.4777	0.5724	0.5134
Bible	Page	GPT-4o	Basic Vector	36.22	0.4000	0.4498	0.4171
Bible	Page	GPT-4o	Hierarchical Hybrid	37.40	0.3945	0.4542	0.4166
Old Test.	Image	GPT-4.1	Basic Vector	40.67	0.4603	0.5310	0.4854
Old Test.	Image	GPT-4.1	Hierarchical Hybrid	40.85	0.4454	0.5306	0.4767
Old Test.	Image	GPT-4o	Basic Vector	36.74	0.3987	0.4612	0.4201
Old Test.	Image	GPT-4o	Hierarchical Hybrid	35.15	0.3771	0.4403	0.4003
Old Test.	**Page**	**GPT-4.1**	**Basic Vector**	**44.23**	**0.5372**	**0.6118**	**0.5633**
Old Test.	Page	GPT-4.1	Hierarchical Hybrid	42.36	0.5106	0.6066	0.5463
Old Test.	Page	GPT-4o	Basic Vector	37.65	0.4449	0.5101	0.4680
Old Test.	Page	GPT-4o	Hierarchical Hybrid	37.89	0.4293	0.5041	0.4570

Within retrieval methods, basic vector search consistently outperformed the hierarchical hybrid search during Phase 1. This was true regardless of prompt, LLM, or view type. The difference was particularly evident when full-page input and GPT-4.1 were used, where vector search reached its highest scores.

The best-performing configuration in Phase 1 used the "Old Testament illustration" prompt, full-page input, GPT-4.1 for description generation, and basic vector search. This configuration achieved a precision of 0.5372, a recall of 0.6118, and an F1 score of 0.5633, with a weighted score of 44.23. This made it the clear choice to carry forward into Phase 2.

The second-strongest configuration—also using full-page input and GPT-4.1 but with the "Bible illustration" prompt—performed only slightly worse, with a precision of 0.4993, recall of 0.5760, and F1 of 0.5265. However, the OT-specific prompt provided a consistent edge across several retrieval setups, suggesting that the increased specificity in the prompt helped the model anchor its descriptions more reliably in the biblical context.

The inclusion of secondary ground truth labels produced only minor adjustments to the results. For the best Phase 1 configuration, the precision and recall values changed only slightly when secondary labels substituted primaries in the applicable sixteen cases. With secondaries included, the best configuration reached a precision of 0.5412, a recall of 0.6046, and an F1 of 0.5628, yielding a weighted score of 44.70. These minimal shifts indicate that the presence of secondary labels—while conceptually important for a small subset of scenes—did not substantially alter the overall performance pattern.

Although Phase 1 successfully identified an effective configuration for medieval manuscript illumination, substantially outperforming both the image-based and BM25 baselines, it underperformed relative to the same methodology applied to early modern woodcuts. Nevertheless, within the manuscript domain itself, the relative performance differences remain clear. The image-only baseline demonstrated that visual similarity alone is insufficient. BM25 search provided a stronger baseline, confirming that many descriptions contained useful lexical cues yet still lacked the semantic depth needed for consistent disambiguation. Against these baselines, the Phase 1 configuration—full-page context, "Old Testament illustration" prompt, GPT-4.1 description generation, and vector retrieval—not only provided the strongest results but also demonstrated that the methodology developed

for the woodcut corpus does transfer to medieval illumination, albeit with reduced absolute performance. Phase 1 thus established a strong baseline and identified the configuration carried forward into the hierarchical expansions of Phase 2.1.

4.3. Phase 2.1 Results: Adding Parent-Level Hierarchy

Phase 2.1 evaluated whether incorporating Iconclass parent hierarchies into the RAG step improved classification performance over the best configuration from Phase 1. As described in the methodology, Phase 2.1 was divided into two experiments. Phase 2.1a tested flat and hierarchical expansions of parent codes arranged purely by Iconclass order; Phase 2.1b repeated the experiment but restored the original top-five vector ranking alongside the expanded parent lists to preserve the retrieval context that had been lost in 2.1a.

4.3.1. Phase 2.1a: Parent Expansion Without Ranking Context

Phase 2.1a showed that providing the LLM with the parent-level structure clearly improved accuracy over Phase 1. Both the flat and hierarchical presentations produced stronger performance, but the hierarchical formatting, using indentation to show structural relationships, yielded the best results of this stage.

Using the primary ground truth only, the flat variant achieved a precision of 0.5653, recall of 0.6192, and an F1 score of 0.5791, with a weighted score of 45.52. When parent codes were arranged hierarchically, the metrics improved further: precision rose to 0.5794, recall to 0.6242, and F1 to 0.5872, with a weighted score of 46.35. These represent the highest scores obtained so far and indicate that parent-level context helps the model to avoid over-specific or structurally inconsistent leaf codes.

Substituting secondary ground truth labels where available yielded similar results: the hierarchical variant again performed best, with a precision of 0.5831, recall of 0.6170, F1 of 0.5860, and weighted score of 46.75. The stability of these results suggests that the addition of secondary ground truth labels did not materially shift the model's overall performance and that the hierarchical representation reliably enhances the model's interpretive accuracy.

Overall, Phase 2.1a demonstrated that hierarchical awareness benefits the model. However, it also revealed a methodological flaw: by sorting parent codes purely numerically, the model lost visibility into the original vector ranking, which had been one of the strengths of the Phase 1 RAG step.

4.3.2. Phase 2.1b: Restoring Ranking Context

Phase 2.1b addressed this problem by supplementing the parent-expanded lists with the original ordering of the top five vector results. This adjustment reintroduced retrieval confidence information that the LLM could use to resolve ambiguous or closely related scenes.

Surprisingly, restoring the ranking context did not lead to improved scores. In fact, both the flat and hierarchical variants produced slightly lower performance than their Phase 2.1a counterparts. Using the primary ground truth, the flat variant reached a precision of 0.5417, recall of 0.6155, and F1 of 0.5671, with a weighted score of 44.55. The hierarchical variant performed marginally better, with precision 0.5476, recall 0.6195, F1 0.5715, and weighted score 44.85, yet still fell short of the 2.1a hierarchical scores.

Secondary ground truth substitution produced the same pattern: the flat approach reached a weighted score of 44.98, and the hierarchical version 45.28, both lower than the corresponding 2.1a results.

The decline in performance suggests that reintroducing the ranking list may have distracted the model or encouraged over-weighting of the initial vector hits, even when parent-level reasoning would have pushed the model toward a more accurate classification.

In other words, the hierarchical parent tree alone provided a clearer and more semantically structured decision space than the combined tree-plus-ranking format.

Given these findings, the best-performing configuration from Phase 2.1 was clearly the hierarchical parent expansion without the ranking list from Phase 2.1a, which improved precision, recall, F1, and weighted scores beyond all earlier configurations. This version served as the foundation for Phase 2.2, where the "None" option and rescue pipeline were implemented.

4.4. Phase 2.2 Results: Introducing "None" and the Rescue Pipeline

Phase 2.2 added the final structural innovation to the pipeline: giving the model permission to answer, "None of the above," followed by a multi-step rescue procedure. This mechanism enabled the system to recover from misleading top-five retrievals and subtree biases that could not be fixed by parent-level reasoning alone. Phase 2.2 was evaluated in two versions: 2.2a and 2.2b.

4.4.1. Phase 2.2a: Initial Rescue Pipeline

Phase 2.2a introduced the core rescue steps: prompting the model to abstain when necessary, re-running vector search with subtree filters (1 vs. 7), re-evaluating parent lists, attempting book-level identification, and finally applying a fallback classification. The goal was not to increase headline metrics but to correct a small subset of mistaken cases identified in Phase 2.1.

Accordingly, the overall scores in Phase 2.2a were slightly lower than the best configuration in Phase 2.1. Using primary ground truths, Phase 2.2a achieved a precision of 0.565, recall of 0.612, F1 of 0.575, and a weighted score of 45.49. These remain close to the Phase 2.1a hierarchical results but do not surpass them. With secondary ground truths substituted, results remained stable: precision 0.568, recall 0.606, F1 0.574, weighted score 45.90.

Although the aggregate metrics dipped slightly, the rescue pipeline still demonstrated value. The weighted scoring distribution shows that full matches remained stable (17.6%), but catastrophic failures (no matches) remained low at 5.6%, and importantly, several edge cases improved. Six scenes improved, none degraded, and 261 remained unchanged. This indicates that the rescue pipeline targeted—and successfully corrected—precisely the problematic cases it was designed for, without destabilizing the broader system.

4.4.2. Phase 2.2b: Refined Rescue Strategy

Phase 2.2b introduced two refinements: removing overly broad parent levels during early rescue stages, and always attempting 7 parent codes before 1 codes, which reflects the overwhelmingly biblical nature of the corpus. These refinements produced clear and consistent gains.

Using primary ground truths, Phase 2.2b achieved a precision of 0.574, recall of 0.630, F1 of 0.591, and weighted score of 46.53. These results exceed both Phase 2.2a and the best configuration from Phase 2.1, making Phase 2.2b the strongest overall configuration. With secondary ground truths included, the pattern was the same: precision 0.579, recall 0.624, F1 0.591, weighted score 46.91.

The match-type distribution improves subtly but meaningfully. Full matches rise to 18.0%, partial matches remain stable, and crucially, no-match outcomes fall to only 3.4%, nearly half the rate of Phase 2.2a. As in 2.2a, six scenes improved, none degraded, and 261 remained unchanged, indicating that improvements came specifically from the cases where subtree imbalance or misleading retrieval patterns previously caused failure. Taken together, the two phases show that while hierarchical reasoning (Phase 2.1) provides the largest gains, rescue-based corrections (Phase 2.2) offer valuable incremental improvements and significantly reduce unresolvable misclassifications.

4.5. Truncation Results

The final stage of the evaluation applied hierarchical truncation to the best-performing model configuration (Phase 2.2b). Truncation assesses how well the system performs at progressively broader levels of the Iconclass hierarchy by repeatedly removing the deepest level of each predicted code and recalculating the evaluation metrics. This process makes it possible to determine whether incorrect leaf-level predictions nonetheless fall within the correct narrative branch or parent category.

At full depth, the model achieved a precision of 0.5744, recall of 0.6296, and F1 score of 0.5909, with an average code depth of 6.49 levels. After one level of truncation, precision rose substantially to 0.6447, while recall declined moderately to 0.5953, producing an improved F1 of 0.6066 (see Table 2). This behavior mirrors the pattern observed in the earlier woodcut study: trimming the most specific level tended to eliminate many of the model's fine-grained errors while preserving the broader correctness of its classifications. The average remaining depth dropped to 5.49 levels, showing that even after truncation, the predictions remained quite specific.

Table 2. Results from Phase 2 after truncation.

Levels Truncated	Precision	Recall	F1 Score	Avg. Levels
0	**0.5744**	**0.6296**	**0.5909**	**6.49**
1	**0.6447**	**0.5953**	**0.6066**	**5.49**
2	**0.7269**	**0.5469**	**0.6080**	**4.49**
3	**0.8123**	**0.4775**	**0.5797**	**3.49**

With two levels removed, the trend continued. Precision increased to 0.7269, while recall decreased to 0.5469, yielding an F1 of 0.6080—the highest F1 score of the entire truncation curve. This plateau suggests that many incorrect predictions differ from the ground truth only at the lowest one or two levels and that the model's broader hierarchical placement is often sound even when its deepest commitments are wrong. The average remaining depth at this stage was 4.49, still within the range of meaningfully specific Iconclass categories.

Beyond this point, further truncation continued to raise precision but reduced recall more sharply. The truncation results paint a clear picture of the model's hierarchical competence. The model often misidentifies the most specific Iconclass leaf but tends to choose codes that are directionally correct within the larger narrative structure. The strongest balance between specificity and accuracy occurs at one or two levels of truncation, where precision is substantially improved and recall remains reasonably high. This suggests that for applications where broad thematic classification is more important than precise leaf-level coding—such as large-scale corpus exploration, clustering, or thematic browsing—the system performs at a level that may already be practically useful.

The truncation analysis thus confirms that while the system is capable of reliably locating scenes within the correct biblical branch, fine-grained specificity remains the most challenging aspect of automated Iconclass assignment.

5. Discussion

The goal of this study was to determine whether a multimodal RAG pipeline originally developed for early modern woodcuts could be effectively transferred to the fundamentally different visual world of medieval manuscript illumination. The results demonstrate that such a transfer is possible, but they also reveal important differences in how contemporary AI models process these two image traditions. Performance across most metrics was

lower for the *Wenzelsbibel* than for the woodcuts, even after substantial adaptation of the methodology. At the same time, the model's behavior suggests that certain forms of hierarchical and contextual reasoning generalize well across media, confirming that AI-assisted iconographic classification can be meaningfully extended into manuscript studies with appropriate caution.

One of the most striking contrasts with the earlier work lies in the relative benefit of full-page input. In the woodcut study, the difference between illustration-only and full-page images was dramatic: full-page input raised precision and recall substantially, with the best RAG-vector configuration achieving an F1 score above 0.82—significantly higher than any manuscript configuration. Printed pages contain numerous features that are immediately legible to modern AI systems: standardized headings, familiar typographic conventions, uniform typefaces, and consistent chapter structures. The *Wenzelsbibel*, by contrast, presents a far more heterogeneous page environment. Despite this, full-page input still consistently outperformed illustrations alone in Phase 1, though the margin of advantage was considerably narrower than in the woodcut corpus. This result suggests that modern multimodal models are surprisingly capable readers of medieval page architecture, but also that manuscript pages provide less consistently actionable information than early modern printed books.

The most substantial performance gains came in Phase 2.1, where hierarchical parent codes were introduced into the RAG step. Here, the model's behavior was clear: structured hierarchical information matters more than the probabilistic ranking of the top-five vector results. When the parent codes were presented in a clean hierarchical format—with indentation and arrows to indicate structure—the model performed better than when the ranking list was reintroduced. This aligns with a growing body of evidence that LLMs excel at symbolic and logical reasoning when given well-organized conceptual structures, such as hierarchical lists (Tan et al. 2024). In this study, the model repeatedly demonstrated an ability to navigate the Iconclass hierarchy in a top–down manner, using the structure itself to decide which branches were plausible and which specific leaves were semantically inconsistent with the description. The success of this stage suggests that the semantic scaffolding provided by Iconclass is something that contemporary language models can use effectively, even across different artistic media.

In this context, the introduction of the rescue pipeline in Phase 2.2 had a more modest effect. The rescue logic, designed to catch cases where vector results returned irrelevant results, did correct a small number of systematic failures. Yet the overall gains in precision, recall, and weighted score were limited. Phase 2.2a produced performance slightly below the best Phase 2.1 model, and Phase 2.2b introduced only incremental improvements. In total, only six scenes improved in either version, and none degraded. This raises a fair question: was the rescue pipeline worth the additional complexity? From a purely quantitative perspective, its contributions were small, but that could be due to the limitations of this sample corpus. But from a qualitative standpoint, the rescue steps served a valuable function: they corrected the kinds of errors that are most disruptive to cataloging work; cases where the model "falls out of the biblical subtree entirely" or fails to locate even a broadly correct classification. In a real-world annotation workflow, preventing these failures may be more important than raising average scores. The modest quantitative effect, therefore, reflects not the inadequacy of the rescue pipeline but the fact that the core parent-hierarchy method was already performing well.

The distinction between primary and secondary ground truths also warrants discussion. Only sixteen scenes carried secondary codes, and the inclusion of such a small set of these labels had negligible effects on aggregate metrics. This is as expected: the secondary labels are not alternate interpretations but structural corrections, acknowledging

that Iconclass assumes a biblical order that manuscript illumination does not always follow. The model's ability to select secondary codes in several cases reinforces the point that hierarchical matching captures more than exact-match metrics reveal. The model often recognized the precise narrative moment even when Iconclass structurally linked it to a different biblical book. The small numerical effect of secondary labels does not, therefore, diminish their interpretive importance. Instead, it highlights the limits of rigid hierarchical vocabularies when applied to medieval narrative structures. However, it must be noted that Iconclass allows for the insertion of new classifications, but such work was outside the scope of this study.

Finally, the truncation analysis reveals a nuanced pattern: the model's strongest performance occurs not at full depth but with one or two levels removed. This mirrors the earlier woodcut findings, though at lower absolute values. The model frequently misidentifies the most specific Iconclass leaves, yet consistently locates scenes within the correct narrative branch. This suggests that full granularity remains challenging, but broader-level reasoning is robust. For many research purposes, especially exploratory browsing or thematic grouping, such broader correctness is often more valuable than fine-grained precision. Overall, these findings demonstrate both the promise and the limitations of applying AI-based iconographic classification to medieval manuscripts.

6. Conclusions

This study shows that a multimodal RAG pipeline developed for early modern woodcuts can be meaningfully extended to medieval illumination, provided that models receive sufficient page-level context and structured iconographic guidance. While classification accuracy in the *Wenzelsbibel* dataset lags behind woodcut benchmarks, the pipeline consistently recovered correct narrative domains and leveraged hierarchical Iconclass information to stabilize reasoning across visually variable scenes. These findings position multimodal LLMs as promising assistants for manuscript studies, capable of supporting large-scale iconographic analysis while remaining sensitive to the methodological constraints of the field. In addition to accelerating metadata production, the pipeline provides a structured way for historians and art historians to explore large image corpora and compare formal motifs across scenes.

At the same time, the lower performance compared to the woodcut study underscores the distinctive challenges posed by medieval imagery. The model's errors reveal the gap between Iconclass's hierarchy and medieval visual practice, particularly in scenes where a manuscript retells an episode from an earlier biblical book. The modest improvements gained through the rescue pipeline confirm that the system was already performing well at a structural level, but they also highlight a long-term need for approaches that support multidirectional exploration rather than simple top-down or bottom-up traversal within the Iconclass hierarchy.

Several directions now emerge for future enhancement of the pipeline, including the deeper integration of textual metadata. The *Wenzelsbibel* project already contains rich description layers: image captions, editorial illustration notes, and TEI-encoded diplomatic and normalized transcriptions. None of these textual assets was used in the present study. Incorporating them into the RAG step could significantly strengthen description generation, disambiguate visually ambiguous episodes, and anchor the model more firmly in the narrative sequence of the biblical text. For large-scale digital editions, this opens the possibility of pipelines that interleave image understanding with text understanding.

Beyond the *Wenzelsbibel*, the approach evaluated here is well-suited for application to other image corpora where visual interpretation is closely tied to contextual metadata. Collections with rich descriptive layers, such as captions, cataloging notes, or encoded

transcriptions, offer particularly promising settings, as these materials could be integrated directly into the retrieval and reasoning steps of the pipeline. The methodology may also be extended to other forms of visual culture, including paintings, sculpture, or stained glass. Finally, the approach could prove valuable for institutional holdings that possess descriptive metadata but lack access to a standardized iconographic vocabulary comparable to Iconclass, offering a flexible framework for supporting classification, discovery, and comparative analysis across diverse domains.

Another promising direction lies in exploring GraphRAG or similar knowledge-graph-driven retrieval frameworks. Iconclass is inherently graph-like: its categories form parent–child hierarchies, parallel branches, and cross-domain references. In this study, when the model followed the wrong branch, the rescue pipeline could bring it back up the tree but could not guide it laterally toward structurally adjacent branches. GraphRAG offers a mechanism for moving "sideways" as well as "up" and "down," allowing the system to navigate biblical, thematic, and conceptual relationships more flexibly. This could help resolve cases where the model recognized the correct general theme but was stranded in the wrong subtree due to misleading vector similarities or incomplete parent lists.

Finally, the study suggests that Iconclass itself—while powerful—is not always ideally aligned with manuscript classification practices. The logic of Iconclass treats narrative, codicological, and topical dimensions as nested levels within a single hierarchy. Medieval imagery often divides these dimensions differently. One potential improvement would be to experiment with using Iconclass more like FAST subject headings, where multiple headings can be assigned: one for the biblical book (codicological location), one for the narrative episode (topical content), and one for more abstract themes (sacrifice, covenant, lawgiving, etc.). Such an approach would better capture the layered interpretive nature of manuscript illumination and allow AI systems to treat biblical book and narrative identity as related but not structurally fused categories. It would also reduce the penalty imposed by cases where the structurally "correct" Iconclass code does not reflect the manuscript's actual book context.

Looking forward, these directions point toward a future in which AI systems operate not as detached classifiers but as integrated agents embedded within the complex networks of textual, visual, and structural information that define medieval manuscripts. Rather than replacing scholarly judgment, such systems can augment it, scaling routine classification tasks, highlighting ambiguous or inconsistent cases for expert review, and enabling new forms of corpus-level analysis. As historians grapple with the implications of AI in research and writing, this study suggests a model of collaboration in which AI serves as a structured, reasoning-enabled assistant. It navigates hierarchies, retrieves knowledge, proposes hypotheses, and handles the repetitive aspects of data annotation, while human expertise remains essential in shaping interpretations, refining hierarchies, and deciding what constitutes meaningful iconographic truth.

In this sense, the *Wenzelsbibel* study contributes not only to the domain of automated iconography but also to broader methodological conversations in the humanities. It demonstrates both the power and the limits of contemporary AI in handling historical materials and illustrates the kinds of infrastructural, ontological, and collaborative innovations needed to make these tools fully effective in scholarly practice. As digital editions continue to expand and repositories grow in scale and complexity, systems that combine hierarchical reasoning, flexible retrieval, and rich metadata integration will play an increasingly central role in shaping the future of art-historical and textual scholarship.

Author Contributions: Conceptualization, D.B.T.; methodology, D.B.T.; software, D.B.T.; validation, D.B.T. and J.H.; formal analysis, D.B.T.; investigation, D.B.T.; resources, J.H.; data curation, J.H.; writing—original draft preparation, D.B.T.; writing—review and editing, D.B.T. and J.H.; visualiza-

tion, D.B.T.; supervision, D.B.T.; project administration, D.B.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Taighde Éireann—Research Ireland, grant number 21/PATH-A/9655a, as part of the "Visualizing Faith: Print, Piety and Propaganda" project at University College Dublin. The *Wenzelsbibel* project was conducted in partnership with the University of Salzburg and the Austrian National Library and was funded by *Land Salzburg*.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The original data presented in this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.18489317 (accessed on 14 February 2025) under a CC BY 4.0 license. The underlying manuscript images and metadata were obtained from *Die Wenzelsbibel—Digitale Edition und Analyse* and are publicly accessible at https://edition.onb.ac.at/wenzelsbibel (accessed on 14 February 2025).

Acknowledgments: This work represents a collaboration between human expertise and AI assistance. The AI (GPT-5.1) functioned as a coding and research assistant and drafting tool, while the authors maintained responsibility for content validity, pedagogical appropriateness, and final editorial decisions. All information is grounded in peer-reviewed literature and institutional policies, with appropriate citations provided throughout. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Banar, Nikolay, Walter Daelemans, and Mike Kestemont. 2021. Multi-Modal Label Retrieval for the Visual Arts: The Case of Iconclass. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence.*. Vienna: SCITEPRESS—Science and Technology Publications, pp. 622–29.

Banar, Nikolay, Walter Daelemans, and Mike Kestemont. 2023. Transfer Learning for the Visual Arts: The Multi-Modal Retrieval of Iconclass Codes. *Journal on Computing and Cultural Heritage* 16: 32. [CrossRef]

Bergel, Giles, Alexandra Franklin, Michael Heaney, Relja Arandjelovic, Andrew Zisserman, and Donata Funke. 2013. Content-Based Image-Recognition on Printed Broadside Ballads: The Bodleian Libraries' ImageMatch Tool. IFLA WLIC 2013—Singapore—Future Libraries: Infinite Possibilities. Available online: https://repository.ifla.org/handle/20.500.14598/5168 (accessed on 14 February 2025).

Brandhorst, Hans, and Etienne Posthumus. Iconclass Browser. Available online: https://iconclass.org/ (accessed on 14 February 2025).

Hlaváčková, Hana. 2001. Old Testament Scenes in the Bible of King Wenceslas IV. In *The Old Testament as Inspiration in Culture: International Academic Symposium, Prague, September 1995*. Třebenice: Mlýn, pp. 132–39.

Kern, Manfred. 2022–2024. Die Wenzelsbibel—Digitale Edition und Analyse. Ein Kooperationsprojekt des Fachbereichs Germanistik der Universität Salzburg und der Österreichischen Nationalbibliothek. Version 7.0.0, 2025-09-19. Available online: https://edition.onb.ac.at/wenzelsbibel (accessed on 14 December 2025).

Österreichische Nationalbibliothek. n.d. *Wenzelsbibel*. Cod. 2759–2764. Vienna: Österreichische Nationalbibliothek.

Posthumus, Etienne. 2020. Iconclass AI Test Set. Available online: https://iconclass.org/testset/ (accessed on 14 December 2025).

Posthumus, Etienne, Harald Sack, and Hans Brandhorst. 2022. The Art Historian's Bicycle Becomes an E-Bike. Paper presented at 6th Workshop on Computer VISion for ART Analysis In Conjunction with the 2022 European Conference on Computer Vision, Tel Aviv, Israel, October 23.

Reshetnikov, Artem, Maria-Cristina Marinescu, and Joaquim More Lopez. 2022. DEArt: Dataset of European Art. *arXiv* arXiv:2211.01226. [CrossRef]

Robertson, Stephen E., Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*. Gaithersburg: National Institute of Standards and Technology, pp. 109–26.

Santini, Cristian, Etienne Posthumus, Tabea Tietz, Mary Ann Tan, Oleksandra Bruns, and Harald Sack. 2024. Multimodal Search on Iconclass Using Vision-Language Pre-Trained Models. In *Proceedings of the 2023 ACM/IEEE Joint Conference on Digital Libraries*. Santa Fe: IEEE Press, pp. 285–87. [CrossRef]

Springstein, Matthias, Stefanie Schneider, Javad Rahnama, Eyke Hüllermeier, Hubertus Kohle, and Ralph Ewert. 2021. iART: A Search Engine for Art-Historical Images to Support Research in the Humanities. In *Proceedings of the 29th ACM International Conference on Multimedia*. New York: Association for Computing Machinery, pp. 2801–3.

Strezoski, Gjorgji, and Marcel Worring. 2018. OmniArt: A Large-Scale Artistic Benchmark. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14: 1–21. [CrossRef]

Tan, Xiaoyu, Haoyu Wang, Xihe Qiu, Yuan Cheng, Yinghui Xu, Wei Chu, and Yuan Qi. 2024. Struct-X: Enhancing Large Language Models Reasoning with Structured Data. *arXiv* arXiv:2407.12522. [CrossRef]

Theisen, Maria. 2024. The Making of the Wenceslas Bible. With special Consideration of the Theological Concept of its Genesis Initial. In *Luxembourg Court Cultures in the Long Fourteenth Century. Performing Empire, Celebrating Kingship*. Woodbridge: The Boydell Press, pp. 243–81.

Thomas, Drew B. 2022–2026. *Visualizing Faith: Print, Piety and Propaganda*. Dublin: University College Dublin.

Thomas, Drew B. 2026. Automating Iconclass: LLMs and RAG for Large-Scale Classification of Religious Woodcuts. *Digital Culture & Education* 16: 3.

Thomas, Drew B., and Julia Hintersteiner. 2026. Iconclass Annotations and LLM-Generated Descriptions for Selected *Wenzelsbibel* Miniatures. *Zenodo*. [CrossRef]

van de Waal, Henri. 1974–1985. *Iconclass: An Iconographic Classification System*. Edited by Leendert D. Couprie, Rudi H. Fuchs, E. Tholen and G. Vellekoop. Amsterdam: North-Holland Pub. Co.

Wilkinson, Alexander S. 2022. Ornamento Europe: Towards an Atlas of the Visual Geography of the Renaissance Book. In *The Book World of Early Modern Europe*. Leiden: Brill, pp. 547–62.

Wilkinson, Hazel, James Briggs, and Dirk Gorissen. 2021. Computer Vision and the Creation of a Database of Printers' Ornaments. *Digital Humanities Quarterly* 15: 1–9.