Beyond Accuracy: Investigating Vision Model Perception on 19th-Century Decorative Arts

Albina Toumarkine[1], and Chahan Vidal-Gorène[1,2]

[1] École nationale des chartes-PSL University, Paris
[2] Centre Jean Mabillon, Paris

Abstract

This paper examines how pretrained vision models perceive and organize a corpus of 19th-century decorative artefacts and printed materials. Using a zero-shot approach, we combine feature extraction, dimensionality reduction, and clustering to explore how convolutional and transformer architectures respond to historical visual material. Two complementary experiments are presented: the first analyzes corpus-level organization through unsupervised clustering of VGG16 embeddings; the second investigates similarity retrieval from individual queries to compare model interpretability (VGG16, EfficientNet, ViT, DINOv2, and CLIP). By visualizing and aggregating activation maps, we discuss biases in how models attend to shape, ornament, and layout, often emphasizing background contrast or framing over meaningful decorative structure. Rather than measuring accuracy, this study focuses on interpretability and bias, highlighting the challenges of adapting art-historical imagery to contemporary vision pipelines.

Keywords: Islamic ornament, 19th-century ceramics, Unsupervised image analysis, Feature clustering, Interpretability, Vision transformers, Zero-shot learning

1 Introduction

1.1 Background and related work in Art history

During the 19th century, the development of new industrial techniques led to a rapid expansion of applied and decorative arts, stimulating increased artistic, commercial, and institutional engagement with this field. Consequently, this period saw the establishment of new institutions dedicated to supporting the study and promotion of applied arts, reflecting the intersection of artistic practice and industrial production [7; 14]. Early examples include the South Kensington Museum in London (now the Victoria and Albert Museum) and the French Union Centrale des Arts Décoratifs (now the Musée des Arts Décoratifs in Paris).

Alongside this industrial and artistic development, a renewed European fascination with Islamic art began to emerge. Contemporary commentators frequently praised the quality and sophistication of Islamic ornamentation, often juxtaposing it with a perceived artistic decline in the industrializing West [11]. This discourse positioned Islamic art as a valuable source of inspiration for European artists and manufacturers in search of aesthetic alternatives.

This enthusiasm for Islamic models manifested in a wide range of decorative objects that incorporated patterns and motifs drawn from Islamic arts, ranging from precise imitations to more freely adapted designs. Among the various media, ceramics were particularly influenced by Islamic art, often drawing directly on ceramic traditions—most notably Iznik ware, an Ottoman style renowned

Albina Toumarkine, and Chahan Vidal-Gorène. "Beyond Accuracy: Investigating Vision Model Perception on 19th-Century Decorative Arts." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1544–1562. https://doi.org/10.63744/k4Jy3Lr4rdK5.

(a) **(b)** **(c)** **(d)**

Figure 1: 19th-Century European Reinterpretations of Islamic Art

(a) Iznik Dish, c. 1545, earthenware, Paris, Louvre, OA 6643.
(b) Théodore Deck, Dish, 1860-1870, earthenware, Guebwiller, Musée Théodore Deck, THD-993.3.10.
(c) Adalbert de Beaumont, Recueil de dessins pour l'art et l'industrie, Paris, Delâtre, 1859, pl. 54
(d) Théodore Deck, Basin, 1863, earthenware, Paris, Musée des Arts Décoratifs, UC 515.

for its vivid colours, intricate floral and geometric motifs, and refined craftsmanship, mainly produced during the 16th and 17th centuries.

This artistic phenomenon has been widely studied through museum exhibitions, which bring together objects from different periods that are nonetheless closely connected, juxtaposing original artefacts from the Islamic world with the 19th-century decorative objects they inspired [5; 9; 12; 13; 17]. Research arising from these exhibitions often highlights the emergence of new networks in the 19th century, connecting collectors, designers, industrialists, and scholars. It also raises important questions about how visual models circulated at a time when artefacts from the Islamic world were relatively rare and often held in private collections.

Printed materials have been proposed as a potential vector in these transmissions. They encompassed a wide range of formats, including journals specializing in the decorative arts and publications dedicated to Islamic art history. Pattern books, in particular, played a key role in the reproduction and adaptation of ornamental motifs [8]. Over time, designs drawn from Islamic art became increasingly present in these publications, with some volumes devoted entirely to ornamentation in Islamic arts.

Taken together, these developments reflect a proliferation of printed materials that greatly expanded the circulation of ornamental imagery, alongside the rise of serial production methods in the decorative arts. The scale and repetition introduced by these practices were inherently quantitative, revealing the limitations of isolated case studies and underscoring the relevance of large-scale analytical approaches for understanding how artistic forms spread.

To address this question, we combine large-scale image analysis with metadata-driven grouping to identify patterns of influence across visual corpora. A central objective is to assess the dissemination of ceramic designs derived from Islamic arts in 19th-century ceramic production, and to evaluate critically whether their influence has been as significant as traditionally assumed. Within a broader research framework, this project examines the dynamics of artistic transmission during the Industrial Era by modelling visual borrowing and transformation across a corpus of ceramic works and digitized printed materials. In doing so, it raises broader questions about cultural exchange, the diffusion of visual motifs, and the entanglement of art, industry, and printed media in the 19th century.

At a practical level, our aim is to identify connections between objects—here understood as instances of copying or reuse. To this end, we are conducting an extensive similarity search, which will serve as the basis for a network analysis. While human observers can readily perceive such visual similarities across decorative elements, these associations can be more difficult to detect computationally.

This paper presents the first phase of an ongoing research project, during which the question of interpretability emerged as a central concern. As we began applying pretrained computer vision models to an art-historical corpus of 19th-century ceramic artefacts and printed documents, we sought to understand how these models perceive and structure such material in a zero-shot setting. Beyond measuring performance, our aim is to reflect on what these models can meaningfully "see" in this context, as well as on the visual patterns, materials, and stylistic cues that they systematically fail to capture.

1.2 Related work in Computer vision

These initial objectives are grounded in the broader framework of distant viewing approaches [1]. Tracking the circulations of visual forms presents significant methodological challenges that are increasingly addressed through computational means[10].

Recent advances in computer vision have enabled large-scale visual analysis of cultural heritage objects, providing methods directly relevant to this study. For instance, Pondenkandath et al. (2021) developed CNN-based techniques for motif retrieval across different depictions of historical watermarks, demonstrating robust cross-depiction matching [18]. Zhao et al. (2023) applied deep learning to painted pottery, enabling similarity-based retrieval of motifs and visual clustering of stylistic traditions [19]. Elgammal and Saleh (2015) proposed a network model to quantify originality and influence in art history, laying the groundwork for diffusion analysis [6]. Meinecke et al. (2024) addressed the challenge of inconsistent metadata in medieval image corpora using embedding-based visual analytics to support distant viewing approaches[15]. Meyer et al. (2024) combined CLIP with object detection in museum collections to enable visual content-based search, highlighting the potential of vision–language models for multimodal exploration of decorative motifs [16].

Several studies have also explored multimodal and similarity-based approaches to improve interpretability in artwork analysis. CLIP has been applied to the NoisyArt dataset to test zero-shot classification and retrieval capabilities, with GradCAM used to highlight the image regions most strongly associated with textual descriptions [3]. Similarly, the CLIP-Art model fine-tuned CLIP on the iMet collection through contrastive language–image pre-training to address fine-grained artwork classification and retrieval [4]. More recent transformer-based similarity approaches have emphasized attention-based embeddings to model cross-artwork visual affinities, underlining interpretability as a critical concern [2]. While these methods focus primarily on performance and multimodal generalization, our work emphasizes visual interpretability within a historical corpus rather than optimizing cross-domain retrieval accuracy.

Figure 2: Examples from the prints dataset, consisting of digitized 19th-century printed materials.

Figure 3: Examples from the artefacts dataset, comprising photographs of 19th-century ceramic objects.

2 Methodology

2.1 Dataset

The dataset comprises a heterogeneous corpus of historical printed materials and ceramic object photographs, assembled from a variety of institutional and online sources. While some images were gathered through automated harvesting methods (e.g., IIIF manifests), a substantial portion required manual selection to ensure alignment with the research scope. The final cleaned dataset contains 1,932 images for the print collection and 1,758 images for the object photographs. For the experiments conducted here, we worked with a subset of 665 items in total, with 248 images consisting of digitized scans of historical pattern books. For volumes with broader coverage, only the pages or sections directly related to Islamic art were retained. These scans typically feature ornamental motifs in various formats: isolated figures, repeating patterns, or full decorative compositions (Figure 2).

The remaining 381 images are photographs of 19th-century ceramic objects, collected from museum and auction house websites. These images display variation in framing, lighting, and background treatment, and include both wide-angle and close-up views, capturing objects from multiple perspectives (e.g., frontal, oblique, and rear). No normalization or image pre-processing was applied (Figure 3).

Together, the two components form a visual corpus that reflects both historical reproduction (via print) and contemporary documentation (via photography). Each subset poses distinct challenges for computational analysis and requires tailored treatment within the data processing pipeline.

Data Category	Number of Images
Scans of 19th-Century Prints	248
Ceramic Artefact Photographs	381

Table 1: Overview of the two data categories used in the study.

Working with such material exposes a fundamental tension between the art-historical notion of a *corpus* and the computational notion of a *dataset*. In art history, a *corpus* is traditionally an assembled group of works selected to answer particular historical questions and shaped by critical judgment. The main criteria at stake are authorship, provenance, chronology, or context, guided by scholarly aims rather than by scale or standardization. On the other hand, a computer vision *dataset* is constructed to meet computational requirements: it must contain large numbers of consistently formatted and annotated images suitable for algorithmic processing and statistical comparison. Treating art-historical images as datasets therefore raises important challenges: how to preserve meaning, context, and uncertainty within systems designed to favor uniformity and quantification, and, most importantly, how to critically examine the forms of attention, categorization, and visual logic that emerge when computer vision models analyze art in ways that diverge from human

perception and historical understanding.

2.2 Initial strategy and model selection

Given the heterogeneity of the corpus and the absence of annotated data, understanding what a model "sees" in an image is rarely intuitive, and this opacity increases when the model is applied zero-shot to materials that differ substantially from its training data. Before evaluating whether pretrained models can extract art-historically relevant features, it is therefore essential to identify which visual attributes—though irrelevant to the research questions—might nonetheless dominate the model's internal representations. This helps in designing more robust visual pipelines and preprocessing strategies. Thus, our initial objective was to evaluate the behavior of visual models without task-specific adaptation, since the corpus was relatively small and heterogeneous, computational resources were limited, and consistent labels or controlled conditions for meaningful fine-tuning were unavailable.

In a preliminary comparison, we tested several families of models — from multimodal vision–language architectures (OpenAI's CLIP, BLIP, Microsoft's Florence-2) to conventional convolutional backbones (VGG16). These experiments immediately highlighted the importance of interpretability: the groupings produced by vision–language models were difficult to interpret consistently, and Florence-based captioning followed by BERT embedding and PCA projection did not yield clusters that were visually or semantically coherent (see Appendix 7). In contrast, convolutional models such as VGG16 displayed clearer internal structure in PCA space, with clusters that could be related to object shape, decorative density, or layout features. While these groupings were not always the most accurate in a strict quantitative sense, they proved more interpretable than the outputs of vision–language models.

Accordingly, we aimed to assess how pretrained models organize and interpret our dataset under zero-shot conditions and to identify configurations that produce interpretable structures rather than merely optimized ones. To this end, we designed two complementary experiments focusing on interpretability at different scales.

The first examines the global organization of the corpus through unsupervised clustering, revealing the visual regularities that emerge across the dataset as a whole. The second focuses on local relationships between individual images, exploring how specific features drive similarity and visual association within the learned representation space. Together, these experiments form a structured, multi-level approach that allows us to analyze how computer vision models, here VGG16, perceive and structure historical visual material, moving from corpus-level organization to image-level feature associations.

2.3 Experiment 1 – Clustering of visual embeddings

This first experiment applies a standard unsupervised pipeline combining VGG16 feature extraction, PCA dimensionality reduction, and K-Means clustering.

Each image is converted to grayscale, resized to 224 × 224 px, and processed through VGG16 (ImageNet weights, classification layers removed). The resulting embeddings are projected with PCA to identify the main axes of visual variation, then grouped with K-Means to examine whether visually coherent clusters emerge, such as groupings based on shared shape, color, or ornament structure (Figures 8, 9 and 10).

The analysis was conducted separately on the subset of object photographs (K = 7) and on the subset of print scans (K = 10). For the prints, the experiment was repeated with cropped borders for comparison. In both cases, the number of clusters was selected empirically through iterative inspection of clustering outcomes. While this approach provides an exploratory baseline, further

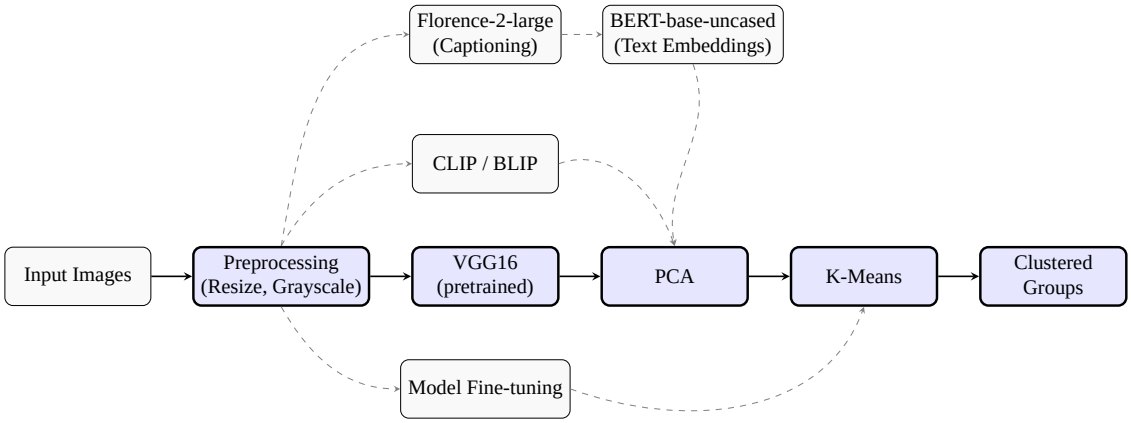

Figure 4: Evaluated pipelines for unsupervised image grouping. All approaches start from a preprocessing stage (resizing, grayscale, channel duplication). The main path (highlighted) uses VGG16 with PCA and K-Means. Alternatives (gray) like Florence-based captioning with BERT, CLIP/BLIP, and fine-tuning were less effective or impractical for our corpus.

refinement—such as using silhouette analysis or hierarchical evaluation metrics—will be required to stabilize K.

To examine the internal coherence of each cluster, we computed a mean feature heatmap by averaging the spatial activations of all cluster members within the last convolutional layer of VGG16. These visualizations help identify which regions of the artefacts most influence cluster assignments, indicating whether the grouping is driven by global form, surface texture, or decorative layout (Figures 11 and 12).

For each image $\mathbf{I}_i$, the pretrained network $f_\theta(\mathbf{I}_i)$ produces a feature vector $\mathbf{z}_i \in \mathbb{R}^d$, which is projected with PCA and grouped using K-Means clustering.

To visualize how clusters are formed, we compute activation maps $H_i = h_\phi(f_\theta, \mathbf{I}_i)$ by averaging the convolutional responses from the last spatial feature layer L of VGG16. Formally, this can be expressed as:

$$\bar{f}_\theta^{(L)}(\mathbf{I}_i) = \frac{1}{C}\sum_{c=1}^{C} f_{\theta,c}^{(L)}(\mathbf{I}_i), \qquad H_i = \frac{\bar{f}_\theta^{(L)}(\mathbf{I}_i)}{\left\|\bar{f}_\theta^{(L)}(\mathbf{I}_i)\right\|},$$

where $f_\theta^{(L)}(\mathbf{I}_i) \in \mathbb{R}^{C\times H_s\times W_s}$ corresponds to the activations from the last convolutional block of VGG16, with C channels and spatial dimensions (H_s, W_s).

For each cluster, a mean heatmap $\bar{H}_k$ is obtained by averaging the activation maps of all its member images:

$$\bar{H}_k = \frac{1}{N_k}\sum_{i\in\mathcal{C}_k} H_i,$$

where $\mathcal{C}_k$ is the set of images belonging to cluster k and $N_k = |\mathcal{C}_k|$. These averaged maps provide an aggregate view of the regions to which the model is most responsive within each group. Random examples from each cluster are displayed alongside their corresponding mean heatmaps to illustrate the characteristic visual patterns captured by the model and to support the interpretation of its attention across clusters.

The combination of K-Means clustering with cluster-level mean heatmaps thus provides an interpretable baseline for observing how a generic CNN organizes form and ornament in 19th-century ceramic artefacts and printed materials.

2.4 Experiment 2 - Similarity-based interpretability

To complement the clustering-based analysis, this second experiment focuses on *similarity retrieval from a single image query*. While Experiment 1 examined the global structure of visual embeddings, here we analyze how models perceive similarity between individual images, an approach comparable to visual comparison in art history.

Objective. We aim to assess the interpretability of model-based similarities and compare their visual focus. Each pretrained model retrieves the most similar artefacts for a given query image, based on its internal representation of form and texture. For each retrieval, we compute a *heatmap* showing which image zones contribute most to similarity, distinguishing meaningful decorative regions (motifs, texture) from irrelevant cues (contours, background, illumination).

Models and embeddings. Five pretrained architectures were evaluated, representing both convolutional and transformer-based paradigms: **CNNs** (VGG16, EfficientNet-B7) and **Transformers** (ViT-Base (patch16-224), DINOv2-Base, and CLIP-ViT-B/32). All were used in a zero-shot setting, operating solely from generic visual priors. Each model transforms an image into a high-dimensional embedding $\mathbf{z}^{(m)} = f_{\theta_m}(\mathbf{I})$, indexed with FAISS using cosine similarity to retrieve the k nearest neighbours (Figures 13 and 14).

Retrieval and interpretability pipeline. For a query $\mathbf{I}_q$, its embedding $\mathbf{z}_q^{(m)}$ is compared to all others in the shared embedding space $\mathcal{Z}$, and the top-k most similar images are returned. An activation map $H_m = g_{\phi_m}(\mathbf{I})$ highlights the regions most responsible for the similarity. In convolutional models, H_m averages final convolutional activations; in transformer models, it corresponds to mean attention weights per patch. The most activated region $\mathbf{I}_{\text{crop}}^{(m)} = \arg\max H_m$ identifies the visual nucleus driving similarity.

Formally, this process can be expressed as:

$$\bar{f}_{\theta_m}^{(L)}(\mathbf{I}) = \frac{1}{C}\sum_{c=1}^{C} f_{\theta_m,c}^{(L)}(\mathbf{I}), \qquad H_m = \frac{\bar{f}_{\theta_m}^{(L)}(\mathbf{I})}{\left\|\bar{f}_{\theta_m}^{(L)}(\mathbf{I})\right\|}, \qquad \mathbf{I}_{\text{crop}}^{(m)} = \arg\max H_m.$$

where f_{θ_m} denotes the pretrained vision model m (e.g., VGG16, EfficientNet, ViT, DINOv2, CLIP), and $f_{\theta_m}^{(L)}(\mathbf{I}) \in \mathbb{R}^{C\times H_s\times W_s}$ corresponds to the activations from its last spatial feature layer L, with C channels and spatial dimensions (H_s, W_s). For convolutional networks, this layer corresponds to the final convolutional block; for transformer-based architectures, it refers to the last attention block.

Dataset-level aggregation. Local activations $H_m(\mathbf{I}_i)$ are aggregated across the corpus to characterize each model's global visual focus:

$$\bar{H}_m = \frac{1}{N}\sum_{i=1}^{N} H_m(\mathbf{I}_i),$$

where N is the total number of images in the dataset. For VGG, multi-layer activations (`block3-5`) are combined and normalized to integrate texture- and form-level features. All activation vectors are then pooled, reduced with UMAP, and clustered using HDBSCAN, producing recurrent motif clusters that capture each model's "visual signature" (see Figure 5).

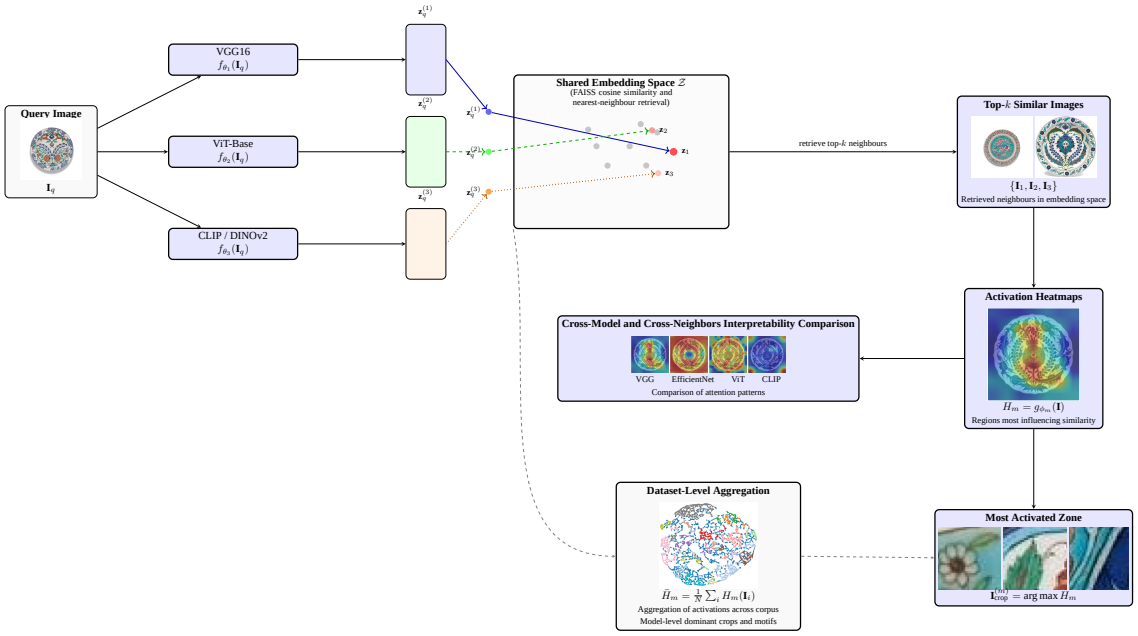

Figure 5: Multi-model similarity interpretability pipeline. Each model retrieves top-k similar images for a query $\mathbf{I}_q$. Heatmaps H_m highlight regions driving similarity, while the crops $\mathbf{I}_{\text{crop}}^{(m)}$ allow cross-model interpretability. At the dataset level, activations from the embedding space are aggregated into $\bar{H}_m$, revealing global attention patterns and dominant motifs across the corpus. For clarity, EfficientNet is not shown in the diagram.

2.5 Results and discussion

Evaluation protocol The pipeline was applied independently to the two subsets of our corpus: one comprising photographs of ceramic objects, and the other consisting of scans of printed materials. The goal in both cases was to organize the images into more coherent subgroups, facilitating further segmentation and cleaning.

Although the broader aim of the project is to investigate visual similarity between printed materials and objects, direct cross-dataset similarity scores between paper-based and ceramic images are generally low, reflecting inherent differences in medium, scale, and texture. By first concentrating on a single modality, we can identify which visual features and patterns the model prioritizes, providing insight to guide the design of subsequent cross-dataset similarity analyses.

Cluster structure (PCA) When clustered via K Means (K=7), the artefact embeddings split into visually coherent groups along two principal axes (see Figure 8). The first principal component (PC1) seems to correlate with object shape: elongated, bottle form vessels exhibit high positive PC1 scores, whereas flat bowls and plates fall at the negative extreme. This suggests that VGG16 filters encode “verticality” versus “flatness” as a primary criterion of variation. The second component (PC2) aligns with decorative complexity: richly ornamented items—with dense floral or geometric motifs—occupy higher PC2 values, while more sparsely decorated or monochrome pieces load lower.

In the PCA plot of the full document images, several distinct visual groupings emerged (see Figure 9). The first principal component (PC1) captures large-scale layout differences: documents with heavy or full-page illustrations tend to lie on one end of the spectrum, while simpler documents—such as those showing isolated motifs or lighter printed layouts—are positioned toward the opposite side. The second component (PC2) appears to reflect a mixture of layout density

and tonal contrast. For instance, documents with tightly packed visual information—such as grid-like arrangements of repeating motifs—cluster along one axis, while documents with broader white space or lower contrast occupy a more dispersed region.

However, closer inspection of the plot suggests that VGG16 may rely heavily on structural or framing features, such as dark zones in the background or shadowed page edges. These features, while visually consistent, are not semantically meaningful within the goals of this project, which seeks to distinguish actual content types (e.g., ornaments vs. object representations vs. architectural plans).

Mean activation maps per cluster The average activation heatmaps reveal a marked difference in the nature of the visual cues driving cluster formation in the two datasets. In the artefacts clusters, activations align with object contours and ornamented surfaces, indicating that clustering is driven by meaningful visual content such as shape and decorative density. In contrast, the document clusters show strong activation concentrated in the corners, suggesting that the model responds primarily to borders and binding shadows. Pages from the same volume frequently fall into the same cluster, not only because of similarities in their visual content, but also due to recurring layout and framing effects.

The impact of this reliance becomes clearer in the cropped-documents PCA–KMeans projection, where such peripheral cues were intentionally removed to evaluate their impact (see Figure 10). Compared to the uncropped version, the overall structure of the PCA projection becomes more compact and internally coherent. Several clusters visible in the uncropped version collapse or reorganize, suggesting a stronger dependence on the central graphic content. Notably, in the cropped version, images representing isolated motifs and those containing continuous ornamental patterns begin to separate more cleanly, suggesting that the removal of background context has allowed the network's representation to focus more on visual differences in the central subject matter. Additionally, certain architectural prints, which often had strong vertical lines or symmetrical structures, become less dominant in the representation space, further supporting the idea that the network was previously influenced by framing rather than semantic content.

Similarity-based interpretability The retrieval results in Figure 5 highlight distinct interpretability patterns across the evaluated models. VGG16 provides the most spatially focused heatmaps, consistently emphasizing the central ornamental medallion and its surrounding radial motifs. The retrieved neighbours closely mirror these structural features, indicating that VGG16 relies heavily on mid-level form and texture. EfficientNet also retrieves artefacts with comparable overall composition, but its activation maps differ in focus: they often place stronger weight on the border of the object and the circular framing elements. In contrast, ViT displays broad and patchy attention spread. The activation patterns appear more evenly distributed across the surface of the object. While the retrieved images generally maintain comparable overall organization, the attention maps show less evidence of consistent focus on specific motifs, implying that the model may be capturing similarity in terms of broader shape or layout. CLIP likewise retrieves artefacts that resemble the query in composition and palette, but its heatmaps more frequently highlight peripheral or low-texture regions. While it is difficult to attribute this behaviour to a single factor, it is consistent with broader representational tendencies observed in multimodal models, where similarity may be influenced jointly by local and global visual cues as well as broader stylistic or contextual associations.

Metrics limitations Our evaluation is qualitative, as the dataset does not include ground truth labels or predefined categories that would allow the use of standard clustering metrics. We therefore assess cluster coherence by manually inspecting grouped images and analyzing their spatial

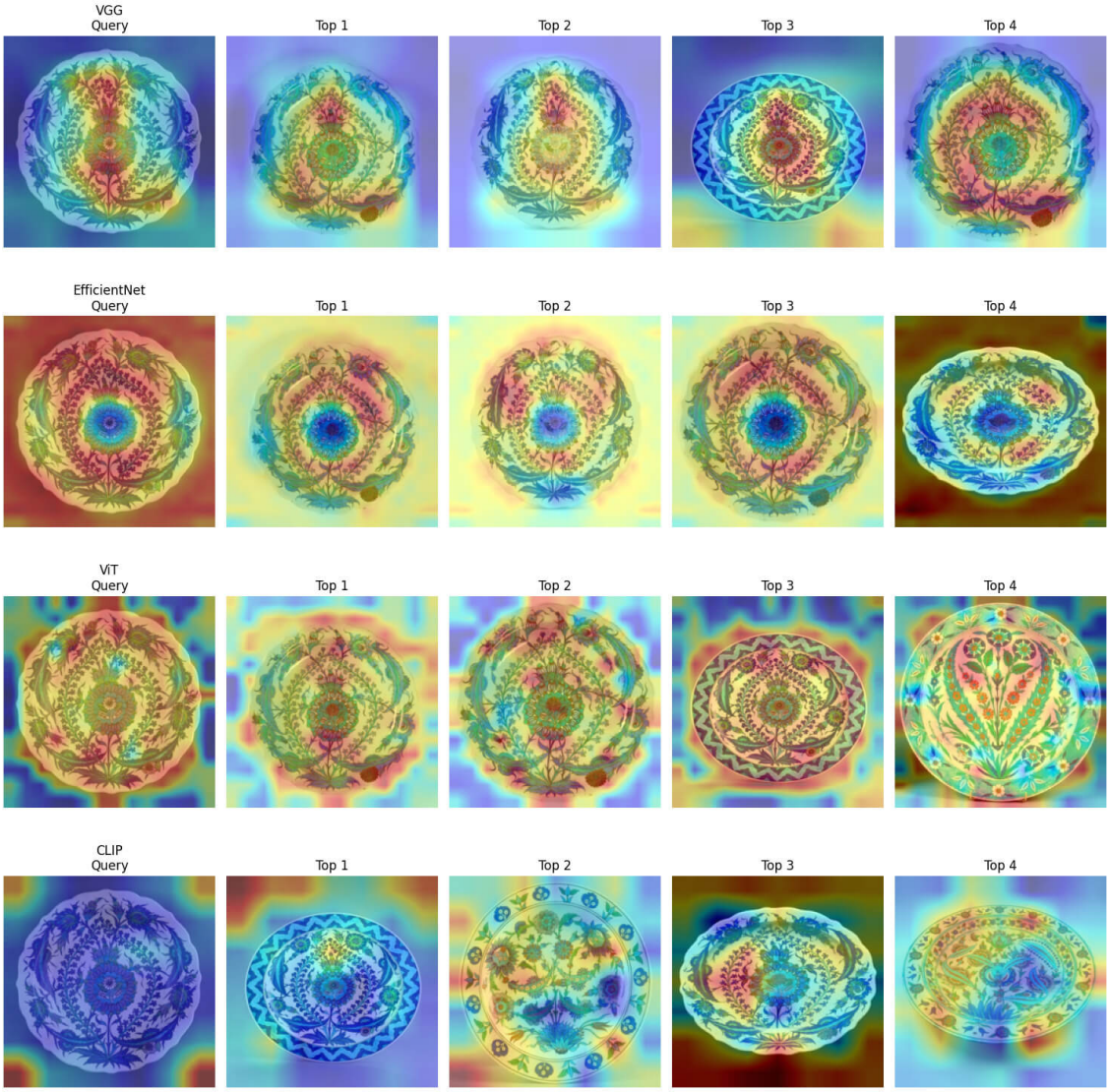

Figure 6: Visualization of Experiment 2 results. For each model, the heatmaps show the query image (left) and its most similar retrieved artefacts. Warmer regions indicate areas most influential in the model's perception of resemblance.

distribution in the PCA projection.

3 Conclusion

This study applies a standard unsupervised pipeline—VGG16 feature extraction, PCA, and K-Means clustering—to photographs of 19th-century artefacts and digitized printed materials, without any task-specific training or fine-tuning. This setup reflects common constraints in digital humanities: limited data, heterogeneous image formats, and scarce resources for model adaptation.

Our observations refine prior insights on multimodal models applied to cultural heritage. While transformer-based pipelines demonstrate remarkable retrieval performance, their attention maps often reflect semantic or stylistic associations rather than strictly visual structures. In contrast, the convolutional hierarchy of VGG16 produces spatially coherent activations that align more closely with human perceptual logic—edges, contours, symmetry, and localized ornament. This contrast

highlights that higher-performing multimodal transformers are not necessarily more interpretable for art-historical purposes. Our results therefore reaffirm the analytical value of simpler convolutional architectures in interpretability-driven research, where the goal is not predictive accuracy but the legibility of visual correspondences.

A central aspect of this work is the adoption of a multi-scale interpretability strategy combining corpus-level grouping with spatial visualizations. At the corpus level, unsupervised clustering reveals how pretrained convolutional features organize the dataset into coherent groups, with mean cluster heatmaps highlighting recurrent regions of model sensitivity that are difficult to discern from individual images. At the image level, single-image heatmaps localize the specific ornamental details driving similarity judgments. This two-step approach forms a practical pipeline linking corpus-level structure to image-level visual evidence.

Even in this zero-shot setting, VGG16 captures some relevant visual properties, including shape, surface texture, and color distribution. While these attributes supported the emergence of partially coherent image clusters, they also revealed structural limitations. In particular, the model often prioritized superficial cues such as framing, background contrast, or page layout—features that do not always align with meaningful distinctions in the context of art historical analysis.

These findings underscore the gap between human visual reasoning and pretrained model behavior. Differences that are significant for researchers—such as decorative type, cultural reference, or motif composition—may be ignored by the model, while insignificant visual regularities are emphasized. Despite this, the interpretability of the representations and heatmaps proved useful for identifying biases and potential avenues for segmentation.

The lack of annotated labels prevents the use of conventional clustering metrics. As such, our evaluation remains qualitative, based on visual inspection of PCA distributions and intra-cluster coherence. This limits the reproducibility and validation of the results and remains a key methodological challenge to be addressed.

In future work, we aim to refine this approach by integrating visual and textual information. Although initial experiments combining Florence-2-generated captions with BERT embeddings did not yield meaningful structures, we believe that caption-based representations may offer complementary signals, particularly when dealing with human descriptions added in catalogs. Exploring such hybrid strategies could improve semantic grouping and help overcome current limitations in working with visually and semantically complex datasets.

Ultimately, while the digitized prints and photographs of ceramic artefacts are studied in our research as a corpus reflecting a single historical phenomenon, our experiments confirm the necessity of maintaining two distinct datasets for computational processing. Although this distinction may appear straightforward in our case, approaching heterogeneous and complex art-historical corpora from a model-interpretability perspective can help reveal subgroups *as perceived by the model*, thereby guiding the steps—such as classification, normalization, or object detection—needed to make an art-historical corpus suitable for computational analysis.

Acknowledgements

This study was conducted as part of the DH master's program at École nationale des chartes–PSL. We thank Léa Saint-Raymond for her valuable feedback and suggestions. It also received support from the PSL Research University's Major Research Program CultureLab, implemented by the ANR (reference ANR-10-IDEX-0001).

Data Availability

Data samples were used exclusively for research purposes and cannot be shared, as they originate from sources such as museums, auction houses, and other cultural institutions.

References

[1] Arnold, Taylor and Tilton, Lauren. “Distant viewing: analyzing large visual corpora”. In: *Digital Scholarship in the Humanities 34*, no. Supplement_1 (Dec. 2019), pp. i3–i16. ISSN: 2055-7671. DOI: 10.1093/llc/fqz013. URL: https://doi.org/10.1093/llc/fqz013 (visited on 07/09/2025).

[2] Asperti, Andrea, Dessì, Leonardo, Tonetti, Maria Chiara, and Wu, Nico. “Does CLIP perceive art the same way we do?” In: *arXiv preprint arXiv:2505.05229* (2025).

[3] Baldrati, Alberto, Bertini, Marco, Uricchio, Tiberio, and Del Bimbo, Alberto. “Exploiting CLIP-based multi-modal approach for artwork classification and retrieval”. In: *International Conference Florence Heri-Tech: The Future of Heritage Science and Technologies*. Springer. 2022, pp. 140–149.

[4] Conde, Marcos V and Turgutlu, Kerem. “Clip-art: Contrastive pre-training for fine-grained art classification”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3956–3960.

[5] Lucien De Guise and William Greenwood, edited by. *Inspired by the East: how the Islamic world influenced Western art [exhibition, British museum, London, 10 October 2019 - 26 January 2020]*. eng. London: The British museum, 2019. ISBN: 978-0-7141-1193-3.

[6] Elgammal, Ahmed and Saleh, Babak. “Quantifying Creativity in Art Networks”. June 2015. DOI: 10.48550/arXiv.1506.00711. URL: http://arxiv.org/abs/1506.00711 (visited on 07/19/2025).

[7] Froissart, Rossella. “« Les collections du Musée des arts décoratifs de Paris : modèles de savoir technique ou objets d’art? »” fr. In: Réunion des musées nationaux, 1994, p. 83. URL: https://amu.hal.science/hal-02338232 (visited on 07/19/2025).

[8] Froissart, Rossella. “« Recueils d’ornements au XIXe siècle : avatars d’un genre épuisé ? », in Ornements : chefs-d’œuvre de la collection Jacques Doucet, sous la dir. de L. Fléjou et M. Decrossas, Paris, INHA, 2015, p. 274-287.” In: *Ornements : chefs-d’œuvre de la collection Jacques Doucet* (Jan. 2015). (Visited on 06/08/2024).

[9] Hagedorn, Annette. *Auf der Suche nach dem neuen Stil: der Einfluss der osmanischen Kunst auf die europäische Keramik im 19. Jahrh.* allemand. Berlin, Allemagne: Staatl. Museen, 1998. ISBN: 978-3-88609-429-5.

[10] Joyeux-Prunel, Béatrice. “Visual Contagions, the Art Historian, and the Digital Strategies to Work on Them”. In: *Artl@s Bulletin 8*, no. 3 (Dec. 2019). ISSN: 2264-2668. URL: https://docs.lib.purdue.edu/artlas/vol8/iss3/8.

[11] Labrusse, Rémi. “Islamic Arts and the Crisis of Representation in Modern Europe”. English. In: *Islamic Arts and the Crisis of Representation in Modern Europe*, ed. by Finbarr Barry Flood and Gülru Necipoğlu. OCLC: 951762809. 2017, pp. 1196–1217. ISBN: 978-1-119-06866-2.

[12] Labrusse, Rémi, décoratifs, Musée des arts, and Louvre, Musée du. *Purs décors ? : arts de l’islam, regards du XIXe siècle, collections des Arts décoratifs [exposition, Paris, Musée des arts décoratifs, 11 octobre 2007-13 janvier 2008]*. fre. Country: FR ill. en noir et en coul., couv. ill. en coul. 29 cm. Bibliogr. p. 352-357. Index. Paris: les Arts décoratifs Musée du Louvre éd, 2007. ISBN: 978-2-916914-02-2 978-2-35031-140-1.

[13] Labrusse, Rémi Directeur de publication, Hellal, Salima Collaborateur, and Beaux-Arts, Musée des. *Islamophilies: l’Europe moderne et les arts de l’Islam*. français. Lyon, France: Musée des Beaux-Arts de Lyon, 2011. ISBN: 978-2-7572-0438-2.

[14] Luneau, Jean-François. “Art et industrie au XIXe siècle : des arts industriels aux industries d’art”. fr. In: *Art & Industrie*. Histoire industrielle et société. Paris: Picard, 2013, pp. 17–24. ISBN: 978-2-7084-0938-5. DOI: 10.3917/pica.stosk.2013.01.0017. (Visited on 05/29/2024).

[15] Meinecke, Christofer, Guéville, Estelle, Wrisley, David Joseph, and Jänicke, Stefan. “Is Medieval Distant Viewing Possible? : Extending and Enriching Annotation of Legacy Image Collections using Visual Analytics”. arXiv:2208.09657 [cs]. Apr. 2024. DOI: 10.48550/arXiv.2208.09657. URL: http://arxiv.org/abs/2208.09657 (visited on 07/19/2025).

[16] Meyer, Louie, Aaen, Johanne Engel, Tranberg, Anitamalina Regitse, Kun, Peter, Freiberger, Matthias, Risi, Sebastian, and Løvlie, Anders Sundnes. “Algorithmic Ways of Seeing: Using Object Detection to Facilitate Art Exploration”. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–18. ISBN: 979-8-4007-0330-0. DOI: 10.1145/3613904.3642157. URL: https://dl.acm.org/doi/10.1145/3613904.3642157 (visited on 07/18/2025).

[17] Hana Nováková, Národní Galerie v Praze, and Zámek Zbraslav, edited by. *The Tulip in the Light of a Crescent Moon: Turkish Ceramics of the 15th to 17th Centuries and their Echoes in Europe ; [october 5, 2003 - february 8, 2004, Collection of Asian Art, Zbraslav Chateau]*. eng. Prague: National Gallery, 2003. ISBN: 978-80-7035-276-2.

[18] Pondenkandath, Vinaychandran, Alberti, Michele, Eichenberger, Nicole, Ingold, Rolf, and Liwicki, Marcus. “Cross-Depicted Historical Motif Categorization and Retrieval with Deep Learning”. eng. In: *Journal of Imaging 6*, no. 7 (July 2020), p. 71. ISSN: 2313-433X. DOI: 10.3390/jimaging6070071.

[19] Zhao, Xiaohan, Shu, Chang, Jiang, Shuping, and Hu, Yutong. “From classification to matching: A CNN-based approach for retrieving painted pottery images”. In: *Digital Applications in Archaeology and Cultural Heritage 29* (June 2023), e00269. ISSN: 2212-0548. DOI: 10.1016/j.daach.2023.e00269. URL: https://www.sciencedirect.com/science/article/pii/S2212054823000140 (visited on 07/19/2025).

A Appendix

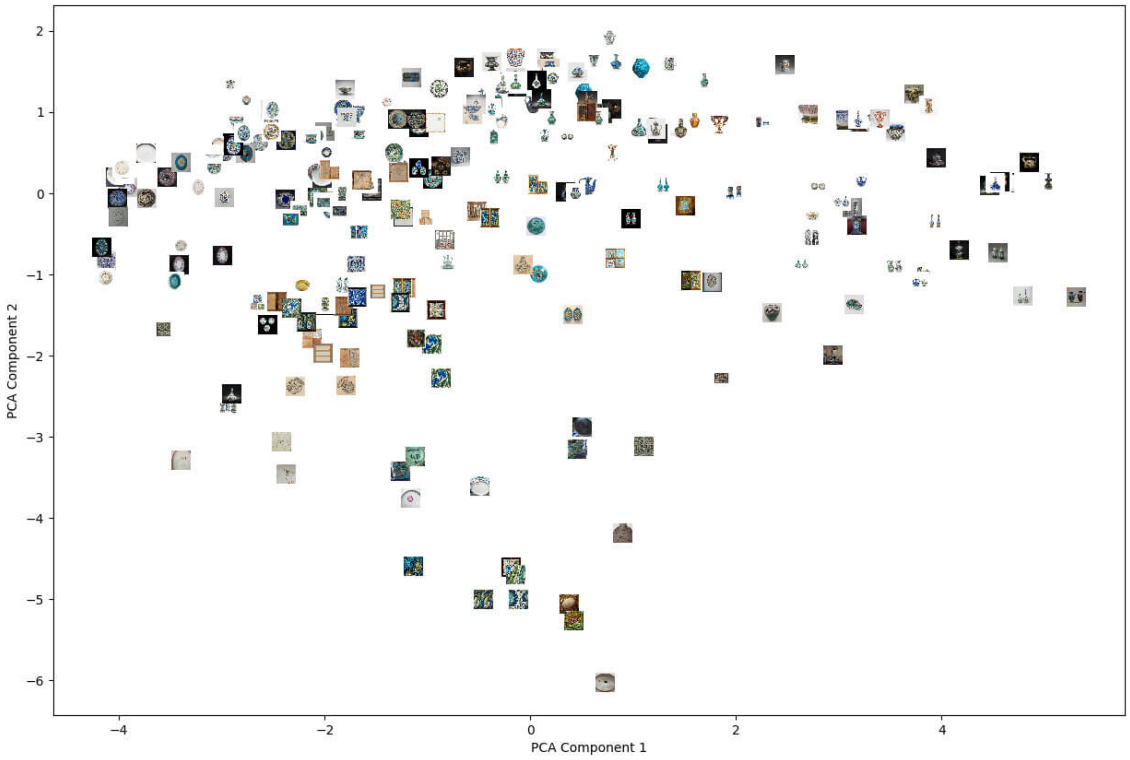

Figure 7: PCA of BERT embeddings from Florence-2-large-generated captions.

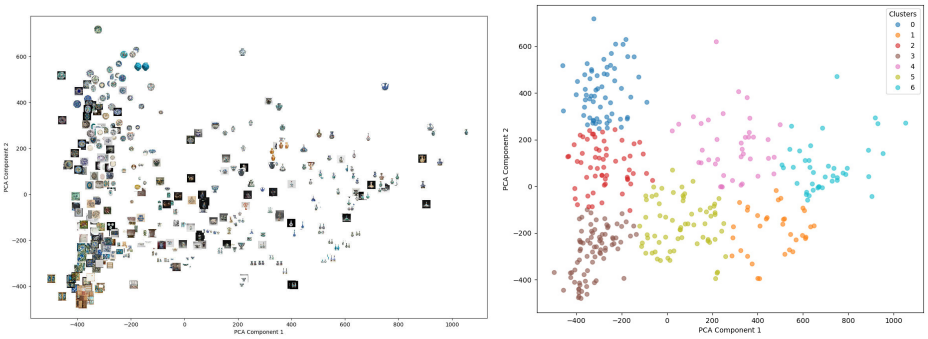

Figure 8: Visual Clustering of the Artefacts Dataset: K-Means Labels and PCA Image Projection

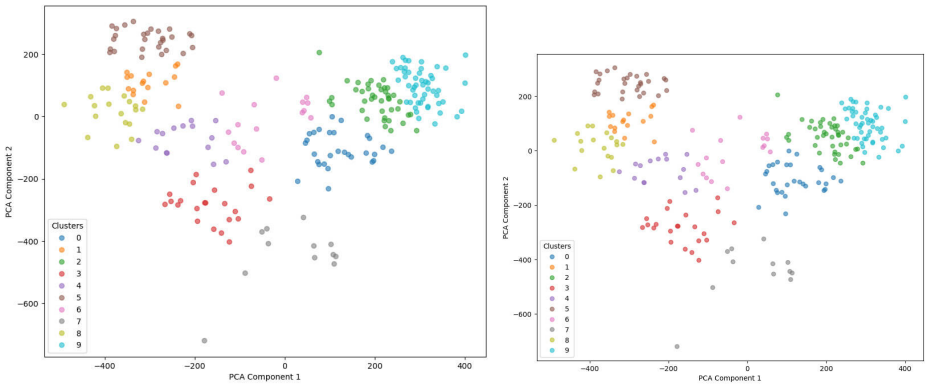

Figure 9: Visual Clustering of the Prints Dataset: K-Means Labels and PCA Image Projection

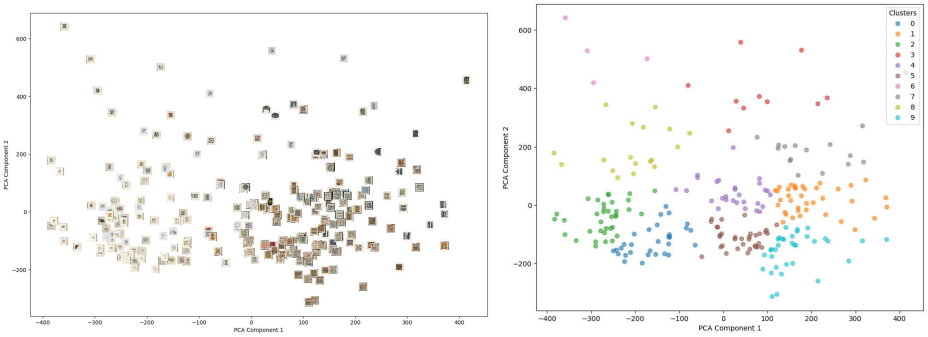

Figure 10: Visual Clustering of the Prints Dataset with Image Cropping: K-Means Labels and PCA Image Projection

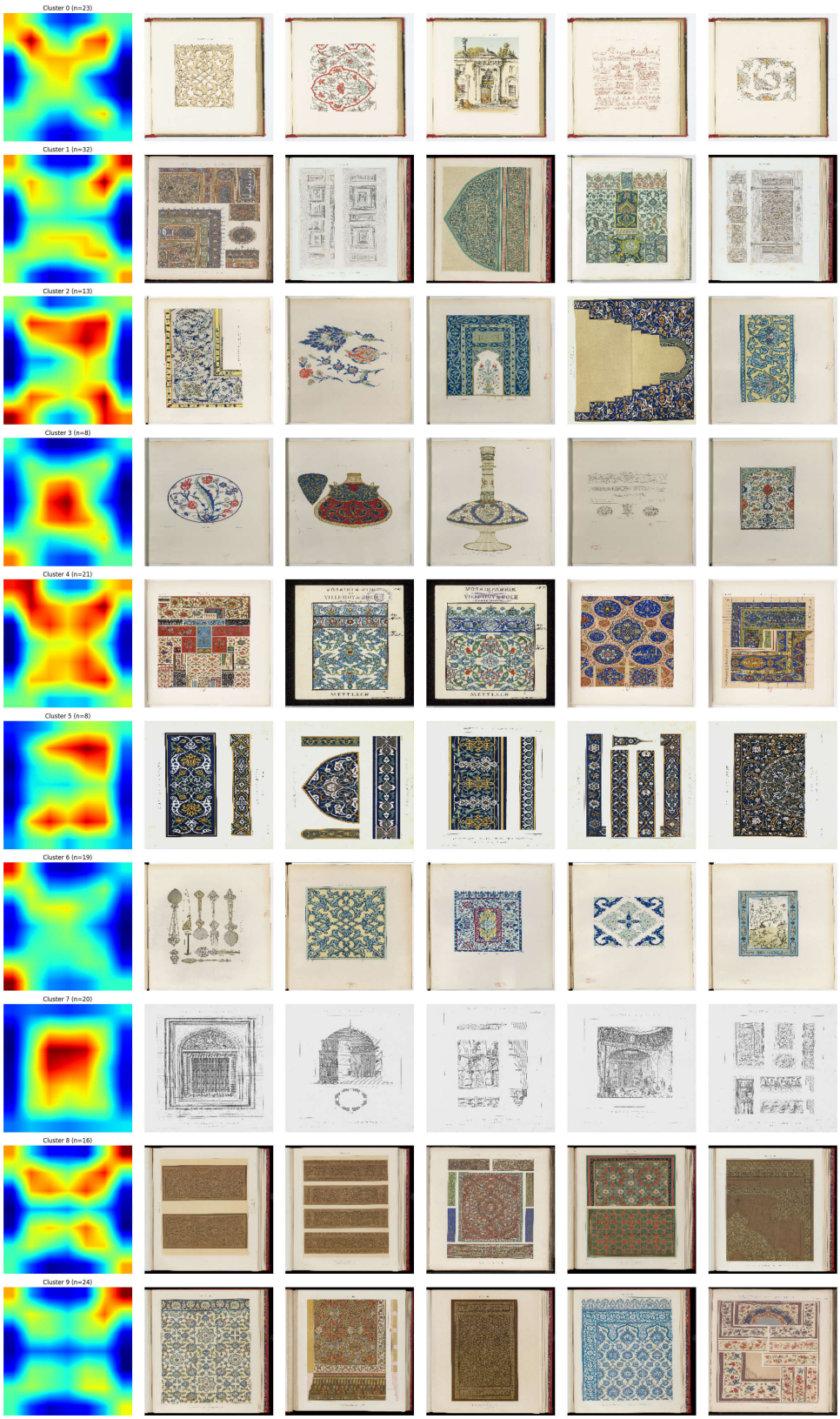

Figure 11: Cluster visualization based on PCA-projected VGG16 features, with mean activation heatmaps computed for each cluster. The first column shows the average heatmap of all items in the cluster, highlighting regions most attended by the model. The following five columns display randomly selected images from the cluster.

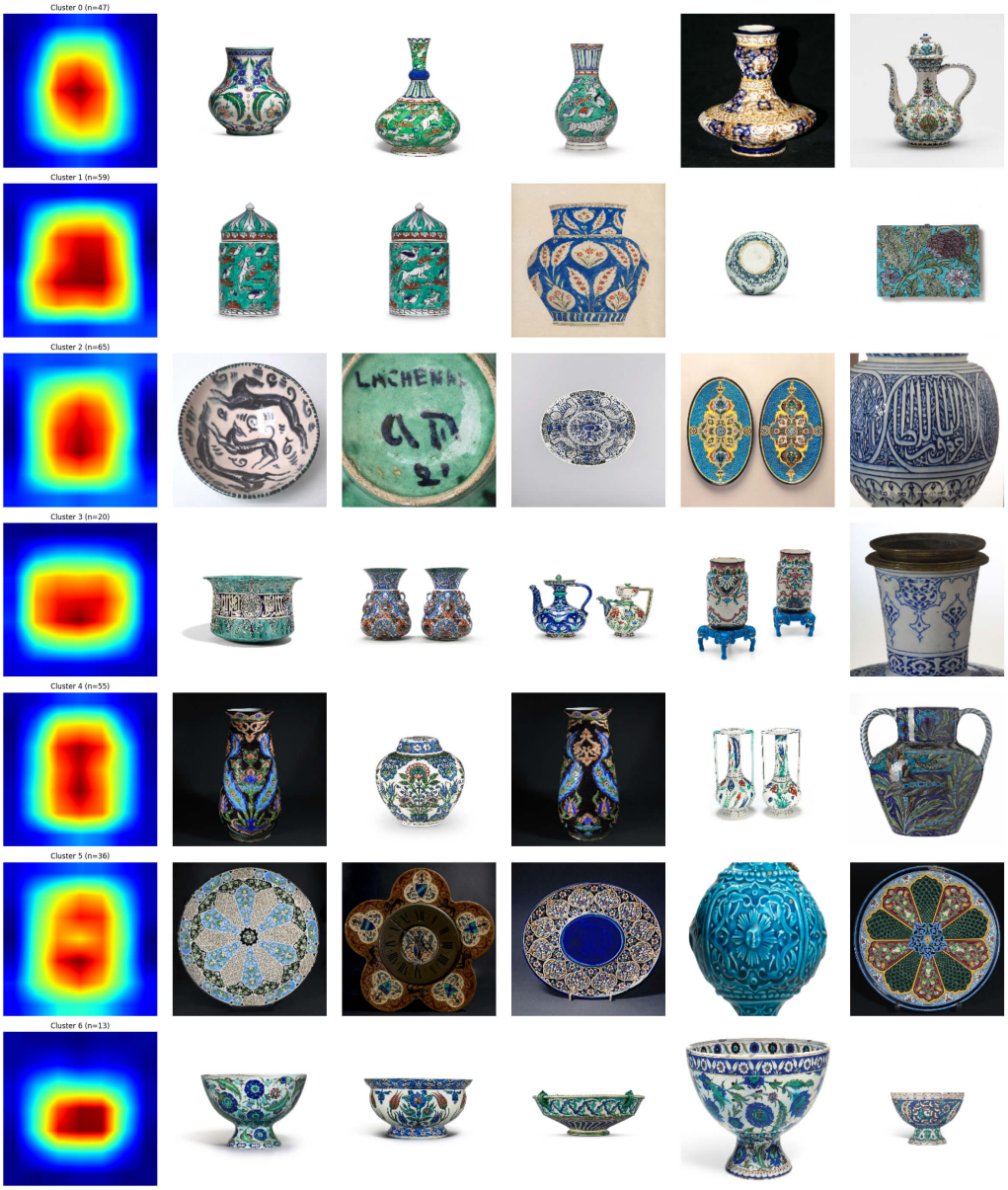

Figure 12: Cluster visualization based on PCA-projected VGG16 features, with mean activation heatmaps computed for each cluster. The first column shows the average heatmap of all items in the cluster, highlighting regions most attended by the model. The following five columns display randomly selected images from the cluster.

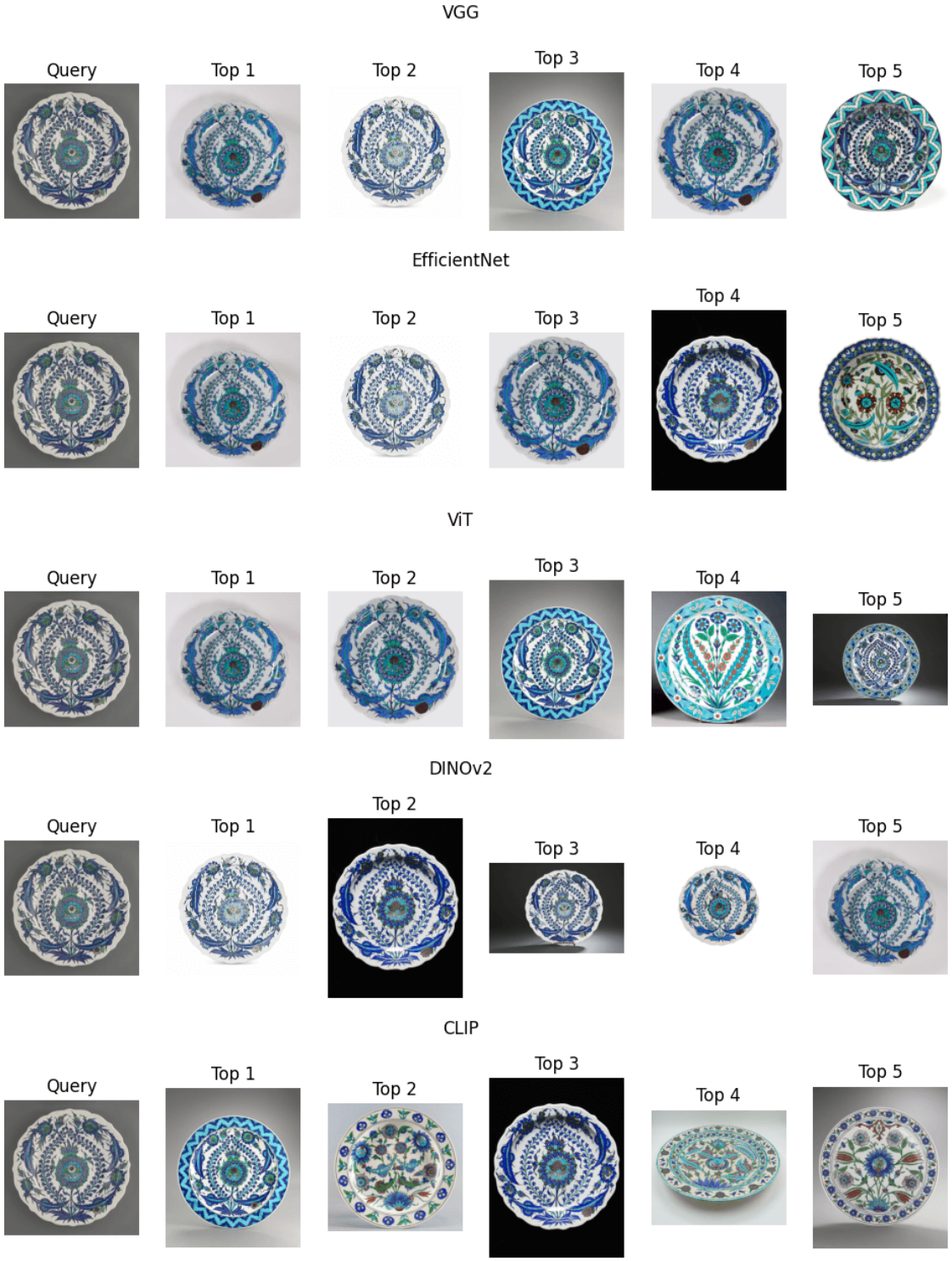

Figure 13: Example of similarity retrieval results for Experiment 2. For each model, the first image (left) is the query, followed by the top-k most similar artefacts retrieved in the embedding space.

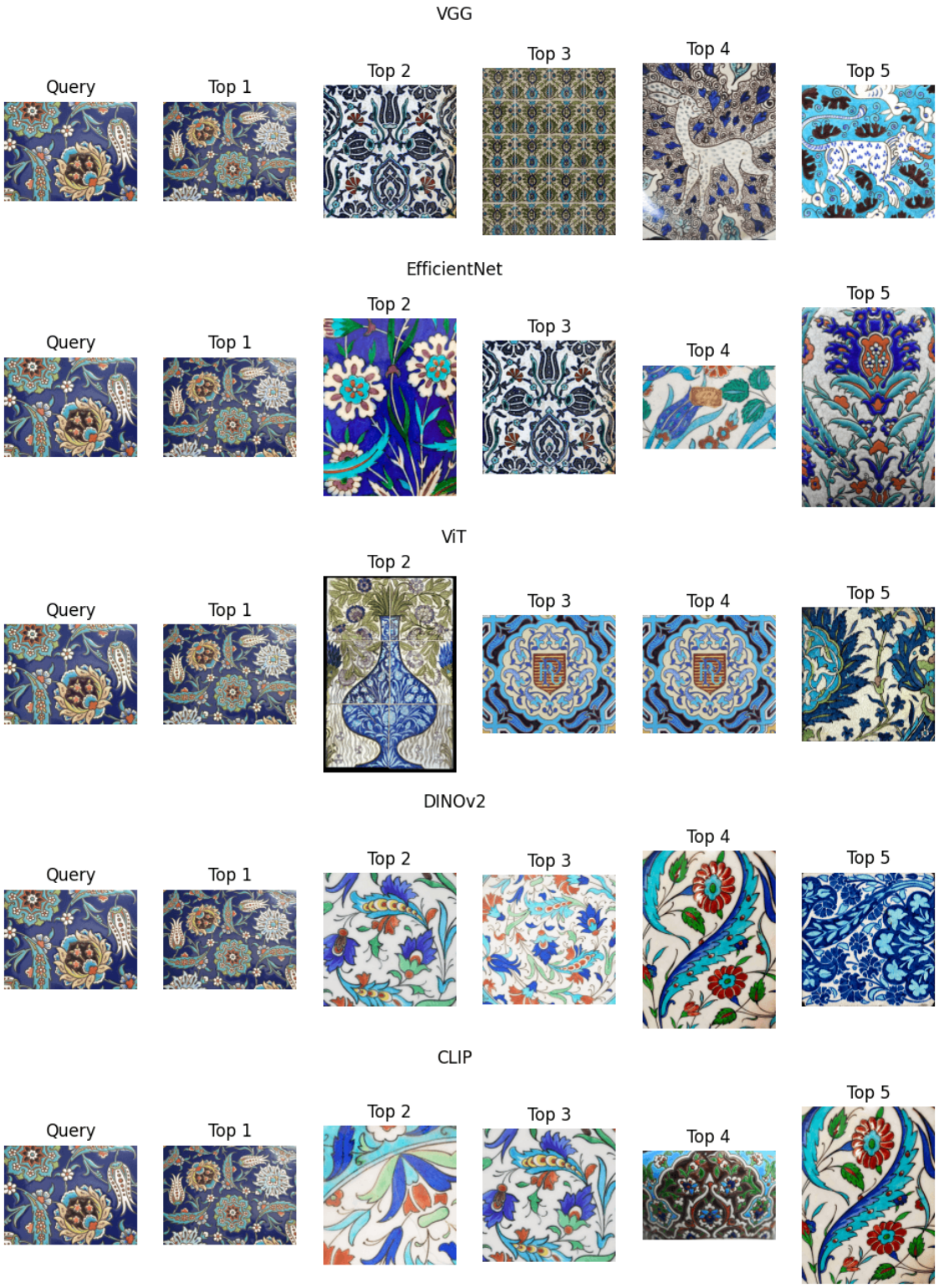

Figure 14: Example of similarity retrieval results for Experiment 2. For each model, the first image (left) is the query, followed by the top-k most similar artefacts retrieved in the embedding space.