

Fachartikel aus:

Gerrit Brüning / Sarah Oberbichler / Cindarella Petz / Pia Schwarz (Hg.): (Generative) KI für Kultur- und Textdaten (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 9). 2026. DOI: [10.17175/sb009](https://doi.org/10.17175/sb009)

Titel:

On the Explainability of Vision-Language Models in Art History

Autor*in:

Stefanie Schneider

Kontakt: stefanie.schneider@uni-marburg.de

Institution: Phillips-Universität Marburg

GND: [1220379301](https://nbn-resolving.org/urn:nbn:de:hbz:5:1-63864-p0011-9) ORCID: [0000-0003-4915-6949](https://orcid.org/0000-0003-4915-6949)

Contribution (CRediT): [Conceptualization](#) | [Investigation](#) | [Methodology](#) | [Visualization](#) | [Writing – original draft](#)

DOI des Beitrags:

[10.17175/sb009_002](https://doi.org/10.17175/sb009_002)

Nachweis im OPAC der Herzog August Bibliothek:

[1982298022](#)

Erstveröffentlichung:

30.09.2026

Lizenz:

Sofern nicht anders angegeben 

Letzte Überprüfung aller Verweise:

27.08.2026

Format:

PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:

[Bildanalyse](#) | [Erklärbare künstliche Intelligenz](#) | [Kunstgeschichte](#)

Empfohlene Zitierweise:

Stefanie Schneider: On the Explainability of Vision-Language Models in Art History. In: Gerrit Brüning / Sarah Oberbichler / Cindarella Petz / Pia Schwarz (Hg.): (Generative) KI für Kultur- und Textdaten (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 9). Wolfenbüttel 2026. 30.09.2026. HTML / XML / PDF. DOI: [10.17175/sb009_002](https://doi.org/10.17175/sb009_002)

Stefanie Schneider

On the Explainability of Vision-Language Models in Art History

Abstract

Vision-Language Models (VLMs) transfer visual and textual data into a shared embedding space. In doing so, they enable a wide range of multimodal tasks, while also raising critical questions about the nature of machine ›understanding‹. In this paper, we examine how *Explainable Artificial Intelligence (XAI)* methods can render the visual reasoning of a VLM – namely, CLIP – legible in art-historical contexts. To this end, we evaluate seven methods, combining zero-shot localization experiments with human interpretability studies. Our results indicate that, while these methods capture some aspects of human interpretation, their effectiveness hinges on the conceptual stability and representational availability of the examined categories.

Vision-Language Models (VLMs) überführen visuelle und textuelle Daten in einen gemeinsamen Einbettungsraum. Dabei ermöglichen sie eine Vielzahl multimodaler Aufgaben, werfen jedoch zugleich kritische Fragen nach der Natur des maschinellen ›Verstehens‹ auf. In diesem Beitrag untersuchen wir, wie Methoden der *Explainable Artificial Intelligence (XAI)* das visuelle Schlussfolgern eines VLM – konkret CLIP – in kunsthistorischen Kontexten nachvollziehbar machen können. Zu diesem Zweck evaluieren wir sieben Methoden, die Zero-Shot-Lokalisierungsexperimente mit Studien zur menschlichen Interpretierbarkeit kombinieren. Unsere Ergebnisse zeigen, dass diese Methoden zwar einige Aspekte menschlicher Interpretation erfassen, ihre Wirksamkeit jedoch von der konzeptuellen Stabilität und der repräsentationalen Verfügbarkeit der untersuchten Kategorien abhängt.

1. Introduction

In recent years, *Vision-Language Models (VLMs)* have become remarkably versatile instruments of analysis. By aligning visual and linguistic information within a shared embedding space, they can perform a wide range of multimodal tasks – from retrieval¹ and captioning² to zero-shot classification.³ However, this versatility has made them the object of sustained criticism: Not only due to the opacity of their internal mechanisms, but also because of the ethical, sociotechnical, and epistemological assumptions encoded in their design. It has been questioned what forms of ›understanding‹ such models enact, how their embeddings reify social hierarchies, and to which extent their apparent generality conceals dependencies on biased, uncured data.⁴ In short, we might ask: What does it mean for a model to *see*?

This question is particularly relevant in fields where visual meaning is historically and semantically dense – where ›objects‹, in the broadest sense, cannot be reduced to mere labels or descriptive tokens. Art history is exemplary in this regard: Here, the visual is not simply perceived, but interpreted through culturally sedimented conventions of style, iconography, and material practice. Nevertheless, models such as *CLIP* (*Contrastive Language-Image Pre-training*)⁵ are now routinely employed ›out of the box‹ for art-historical retrieval and analysis on digital platforms,⁶ often without a clear understanding of which kinds of visual concepts – formal, iconographic, or affective – are encoded in their embeddings.

CLIP is a VLM that aligns images and texts within a shared embedding space. Trained on a large set of image-text pairs scraped from the web, it learns to associate visual and linguistic patterns by grouping similar representations and separating dissimilar ones. Large-scale web-scraped datasets such as LAION-400M⁷,

[1]

[2]

[3]

¹ Cf. Radford et al. 2021.

² Cf. Li et al. 2023.

³ Cf. Zhai et al. 2022.

⁴ Cf. Bhalla et al. 2024; Birhane et al. 2021; Bender et al. 2021.

⁵ Cf. Radford et al. 2021.

⁶ Cf. Springstein et al. 2021; Offert / Bell 2023.

⁷ Cf. Schuhmann et al. 2021.

however, are by no means neutral repositories of visual culture. As Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe demonstrate, such datasets contain structural biases, non-consensual imagery, and stereotypical or pornographic representations that reflect the discriminatory nature of the web.⁸ The epistemic opacity of VLMs thus becomes a methodological issue: How can search results be interpreted when the model itself embeds an unacknowledged theory of vision? CLIP, in particular, epitomizes this multimodal turn: Its embedding space constitutes not only a technical geometry of similarity but what Leonardo Impett and Fabian Offert call a »vector imaginary«⁹ of contemporary visual culture – a statistical condensation of what is collectively pictured and named online. This makes CLIP both uniquely powerful and uniquely problematic for art-historical inquiry. On the one hand, its zero-shot capacity enables the retrieval of artworks that are stylistically or iconographically related without supervision – in other words, it can identify similarities even among categories that it was never explicitly trained on. Yet the same mechanism also perpetuates the omissions of its training corpus, reproducing a visuality that is historically and culturally uneven.

Against this backdrop, we ask a central question: To what extent can *Explainable Artificial Intelligence* (XAI) methods render the visual logic of CLIP legible to human interpreters, thereby strengthening the methodological robustness of VLMs in art-historical contexts? XAI refers to a variety of techniques designed to elucidate model behavior, including post hoc attribution, concept-based analysis, and related methods that indicate which features of an input contribute most strongly to a given output.¹⁰ To explore this question, we comparatively evaluate seven XAI methods spanning three paradigms: (1) Gradient-based methods that backpropagate class-specific gradients into feature maps (Grad-CAM, Grad-CAM++, LayerCAM, and LeGrad);¹¹ (2) score-based, gradient-free methods that measure the influence of image regions on the model-predicted score (ScoreCAM and gScoreCAM);¹² and (3) CLIP-specific approaches that intervene directly in the inference pipeline (CLIP Surgery).¹³ Each of these methods generates a saliency map for a given text prompt that visualizes the contribution of an image region to the model's predicted score (Fig. 1). In this paper, our objective is not just to identify the most effective technique but also to explore the boundaries of explainability under conditions of zero-shot inference and domain transfer. Thus, we included methods that, while no longer state-of-the-art in some cases, remain prevalent in practice, in order to highlight the interpretive and subjective dimensions of ›explainability‹ itself.

[4]

⁸ Cf. Birhane et al. 2021.

⁹ Impett / Offert 2022.

¹⁰ See, e.g., Speith 2022; Saeed / Omlin 2023; Hassija et al. 2024.

¹¹ Cf. Selvaraju et al. 2017; Chattopadhyay et al. 2018; Jiang et al. 2021; Bousselham et al. 2024. CAM denotes Class Activation Mapping.

¹² Cf. Wang et al. 2020; Chen et al. 2022.

¹³ Cf. Li et al. 2025.

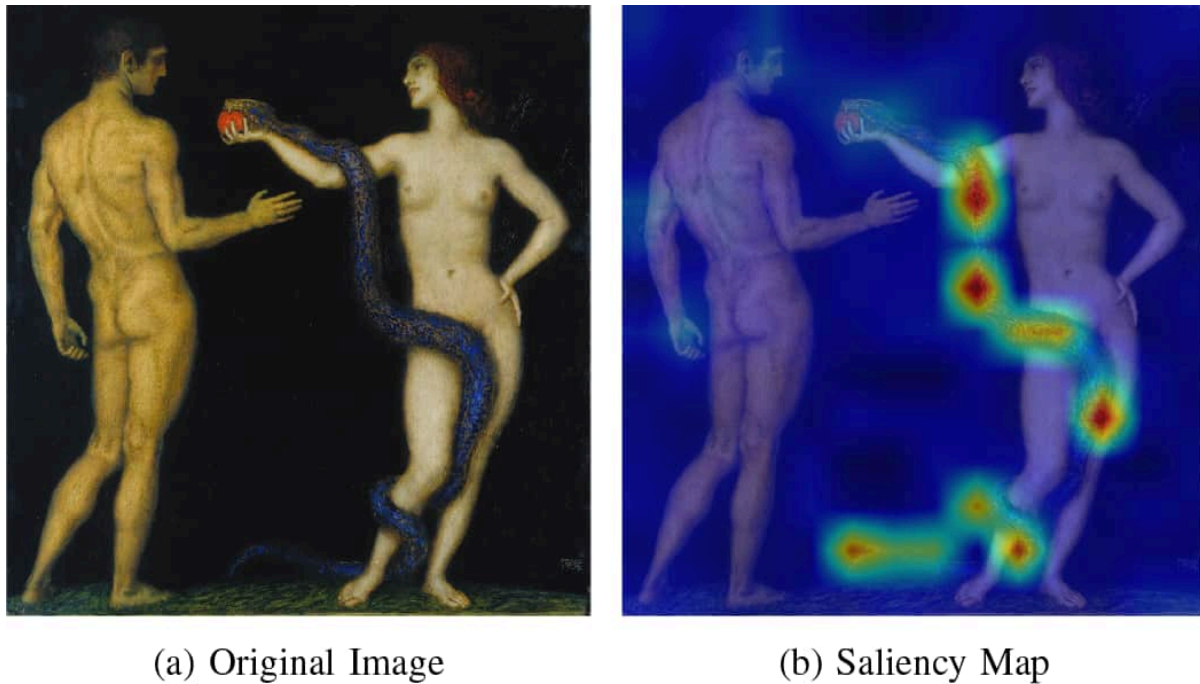


Figure 1: The saliency map highlights, in red, the image regions most strongly associated with the concept of the ›snake‹ in Franz von Stuck's *Adam and Eve* (c. 1920). [Visualization: Stefanie Schneider 2026]

To examine these dimensions systematically, we adopted a two-stage evaluation framework. First, a quantitative case study employed two art-historical datasets – IconArt¹⁴ and ArtDL¹⁵ – to measure the localization accuracy of the methods under zero-shot conditions. Then, in an online survey with participants trained in art history, the interpretability of these same methods was evaluated, situating the previously obtained results within the variability of human visual judgment. Building on the work of Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flávio P. Calmon and Himabindu Lakkaraju, who demonstrate that CLIP embeddings can be decomposed into sparse, interpretable concept spaces,¹⁶ our studies asked whether such interpretability extends to visual explanations: that is, do XAI methods disclose a model's internal conceptual structure – or merely *aestheticize* its opacity? This question is especially pertinent given critiques by Birhane et al. and Emily M. Bender and her Co-Authors Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, who argue that large-scale, web-scraped datasets can perpetuate hegemonic biases under the guise of generality.¹⁷ If explainability techniques cannot illuminate these latent structures, they may reiterate rather than expose the ideological patterns of machine vision.

[5]

From this standpoint, we articulate three research questions: (1) How effectively do XAI methods localize iconographic objects in artworks under zero-shot conditions without fine-tuning? This establishes a baseline: Can methods trained on everyday imagery nonetheless delineate the complex, symbolically charged forms within artworks beyond their training distribution? (2) Does the visual relevance of these maps correspond to human judgments? If saliency maps claim to visualize ›what the model sees‹, their validity must be tested against human perception – specifically against the art-historically informed gaze. (3) Which factors, such as object size and concept abstraction, drive performance differences? Here, we connected measurable attributes with semantic ones, aligning computational error analysis with questions of representation central to art-historical inquiry. By testing how – and where – XAI methods succeed or fail in making CLIP's mechanisms visible, we contributed to a broader methodological debate about how digital art history might

[6]

¹⁴ Cf. Gonthier et al. 2018.

¹⁵ Cf. Milani / Fraternali 2021.

¹⁶ Cf. Bhalla et al. 2024.

¹⁷ Cf. Birhane et al. 2021; Bender et al. 2021.

critically engage with the epistemic structures of machine vision. Our aim, in other words, is to determine not only what VLMs attend to in works of art but examine how their patterns of attention either align – or fail to align – with human interpretive conventions.

The remainder of this paper is structured as follows: In Section 2, we outline the rationale for selecting the examined XAI techniques. Section 3 presents the first case study, which quantitatively evaluates localization accuracy using two art-historical datasets under zero-shot conditions. Section 4 turns to the second case study: an online survey that assesses the interpretability of saliency maps from an art-historical perspective. Section 5 then synthesizes these findings, tracing their methodological implications for explainability, interpretability, and the critical analysis of machine vision within digital art history.

[7]

2. Method Selection

Since the advent of *Artificial Intelligence (AI)*, researchers have sought to make its internal mechanisms intelligible. This imperative has only intensified further with the resurgence of *Deep Neural Networks (DNNs)* in the 2000s – highly non-linear statistical models with millions of parameters. The demand for explanation, however, predates deep learning; it can be traced back to the early attempts to understand expert systems,¹⁸ which, emerging in the 1960s and 1970s, marked one of the first conceptual turns in AI towards the explainability of machine reasoning.¹⁹ Designed to mimic the problem-solving strategies of human specialists within narrowly defined use cases, these systems aimed not only to provide decisions, but also to explain the reasoning behind them.²⁰ Their architecture – typically comprising a knowledge base and an inference engine – explicitly separated domain-specific expertise from domain-independent reasoning procedures.²¹ Although modern DNNs have largely abandoned such symbolic, rule-based representations in favor of statistical learning, the epistemological tension first articulated in the age of expert systems – between algorithmic performance and interpretability – remains a defining problem for contemporary research in XAI.

[8]

In this section, however, we do not intend to provide a historical or taxonomic overview of XAI methods – an exercise already undertaken in recent years from a variety of perspectives.²² Rather, we motivate our decision to focus on a specific class of *perceptive interpretability* methods, i.e., saliency-based visualization methods applied post hoc to a pre-trained model, in our case CLIP. Our goal is not to revisit the broader epistemological debates surrounding explainability, but to examine how visual explanations – those that make a model's internal reasoning visible – operate in the specific and practically relevant context of VLMs. In this context, perceptive interpretability methods, and saliency-based visualizations in particular, occupy a unique position at the interface of human perception and machine inference, as they generate intuitive, human-readable representations of model attention. In doing so, they provide a spatial grammar through which one might then ask where a model ›looks‹ when associating textual prompts with visual content. We therefore selected methods according to three criteria: (1) The method must produce spatially localized, human-inspectable heatmaps over the input image; (2) it must operate post hoc on CLIP – without retraining or architectural modification – to ensure comparability across prompts and datasets; and (3) it must be either widely adopted as a baseline or specifically adapted to CLIP's dual-encoder architecture, making text-image interactions explicit. We excluded approaches that introduce additional hyperparameters, such as CLIP-LIME²³. Likewise, attention-based methods²⁴ were omitted, as their reliance on attention weights has been shown to correlate only weakly with decision relevance.²⁵ Also excluded were methods such as CLIP-Dissect²⁶:

[9]

¹⁸ E.g., Buchanan / Shortliffe (eds.) 1985.

¹⁹ Cf. Coy / Bonsiepen 1989.

²⁰ Cf. Harmon et al. 1989.

²¹ Cf. Puppe 1988.

²² E.g., Speith 2022; Saeed / Omlin 2023; Hassija et al. 2024.

²³ Cf. Kazmierczak et al. 2023.

²⁴ E.g., Abnar / Zuidema 2020.

²⁵ Cf. Jain / Wallace 2019.

Although they yield valuable insights into the representational topology of CLIP, they do not address the specific perceptual question that motivates our study, i.e., where in the image does the model locate the evidence for a given prompt?

Given these constraints, we have identified three methodological paradigms for evaluation, each of which renders the most relevant image regions visible to a model in a different way: *Gradient-based* methods trace how the model's internal signals – its gradients – shift in response to visual stimuli, revealing which regions of an image most strongly determine a given decision; *score-based*, gradient-free methods, by contrast, perturb the input: They mask regions of the image and measure the resulting change in the model's output, inferring from these changes which areas carry the greatest weight; CLIP-specific interventions operate directly on the multimodal architecture, adjusting how CLIP integrates textual and visual information to make visible the relational mechanics between words and image regions. Within the gradient-based family, we include Grad-CAM²⁷ as the canonical baseline, Grad-CAM++²⁸ as its extension to multi-instance scenarios via higher-order gradient weighting, and LayerCAM²⁹ as a refinement that pools activations from intermediate convolutional layers to enhance spatial fidelity. LeGrad³⁰ further optimizes this aggregation process, improving localization accuracy while reducing sensitivity to layer selection. The second paradigm – score-based, gradient-free evaluation – is exemplified by ScoreCAM³¹ and gScoreCAM.³² Both rely on selective occlusion: By systematically masking portions of the image and recording how the model's scores change, they construct saliency maps that recompose these perturbations into a spatial account of the model's visual attention. Finally, CLIP Surgery³³ represents a model-specific intervention at inference time. By reparameterizing the forward pass and adapting Grad-CAM-like mechanisms to CLIP's dual-encoder architecture, it produces activation maps that more explicitly disentangle the textual and visual streams of the model – rendering visible, in effect, how the network's multimodal alignments are spatially instantiated. [10]

However, while such techniques have become – at least to some extent – standard practice within machine vision, their epistemic foundations remain contested. According to Tim Miller, explanations are not static products but dialogical processes that unfold within particular social contexts: They presuppose both an *explainer* and an *explainee*, whose beliefs, expectations, and cognitive biases shape not only the form of an explanation, but also how adequate it is perceived to be.³⁴ As Miller emphasizes, explanations are inherently contrastive (*Why this rather than that?*), selective (*Which of the many possible causes should be prioritized?*), and social (*How should this explanation be shaped for the person to whom it is addressed?*). In this sense, Miller's ›social model of explanation‹ provides a conceptual bridge between the epistemic pragmatics of XAI and the interpretive negotiations that characterize humanistic inquiry. For art-historical applications – where meaning is constructed through dialogue rather than extracted from data – this social-cognitive model is especially pertinent. We, thus, combined quantitative and qualitative experiments to examine how explanation methods function not only as diagnostic instruments but as interfaces between human vision and machine inference. [11]

²⁶ Cf. Oikarinen / Weng 2023.

²⁷ Cf. Selvaraju et al. 2017.

²⁸ Cf. Chattopadhyay et al. 2018.

²⁹ Cf. Jiang et al. 2021.

³⁰ Cf. Bousselham et al. 2024.

³¹ Cf. Wang et al. 2020.

³² Cf. Chen et al. 2022.

³³ Cf. Li et al. 2025.

³⁴ Cf. Miller 2019.

3. Case Study 1

In the first case study, we evaluated the zero-shot localization capabilities of CLIP by examining the afore-introduced XAI techniques using large-scale datasets comprising nearly 2,000 art-historical images in total. In so doing, we determined how effectively each method can identify and delineate objects in domain-specific imagery without any fine-tuning. The aim of this study was to establish a quantitative foundation for assessing how different explainability methods perform in visually and semantically complex cultural datasets, and to identify which approaches transfer most effectively to art-historical imagery. Our analysis focused on comparing these techniques – Grad-CAM, Grad-CAM++, LayerCAM, LeGrad, ScoreCAM, gScoreCAM, and CLIP Surgery – under consistent experimental conditions, thereby delineating the empirical contours against which subsequent interpretive analyses (Section 4) can be situated. [12]

3.1 Pre-processing Steps

For each method, we generated a class-conditional saliency map, thresholded it at τ to obtain a binary mask, and extracted the tightest box around the largest connected component (Fig. 2). These boxes were then compared to ground-truth annotations from the datasets described in Section 3.2. While fixed thresholds (e.g., $\tau = 0.2$) can yield plausible maps, they are notoriously sensitive – to the explainability method, the dataset, and even the target class semantics – leading to unstable conclusions. Following Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata and Hyunjung Shim,³⁵ we therefore adopted a threshold-independent evaluation: For each method, we computed the maximum BoxAcc, which denotes bounding-box localization accuracy at a specified *Intersection-over-Union* (IoU) threshold, over $\tau \in [0.2, 0.9]$ via grid search, ensuring that it was evaluated under the most favorable operating conditions. Unlike Choe et al., who average BoxAcc across IoU thresholds $\delta \in \{0.3, 0.5, 0.7\}$,³⁶ we reported BoxAcc separately at each δ to identify more granular differences in localization quality. Formally: [13]

$$\text{BoxAcc}(\tau, \delta) = \frac{1}{N} \sum_n \mathbb{1}_{\text{IoU}(\text{box}(s(X^{(n)}), \tau), B^{(n)}) \geq \delta}$$

where $s(X^{(n)})$ is the saliency map for image $X^{(n)}$, $\text{box}(s, \tau)$ is the tightest bounding box enclosing the largest-area connected component obtained by thresholding s at τ , and $B^{(n)}$ is the corresponding ground-truth bounding box. In simple terms, BoxAcc indicates whether the highlighted region overlaps sufficiently with the true object location.

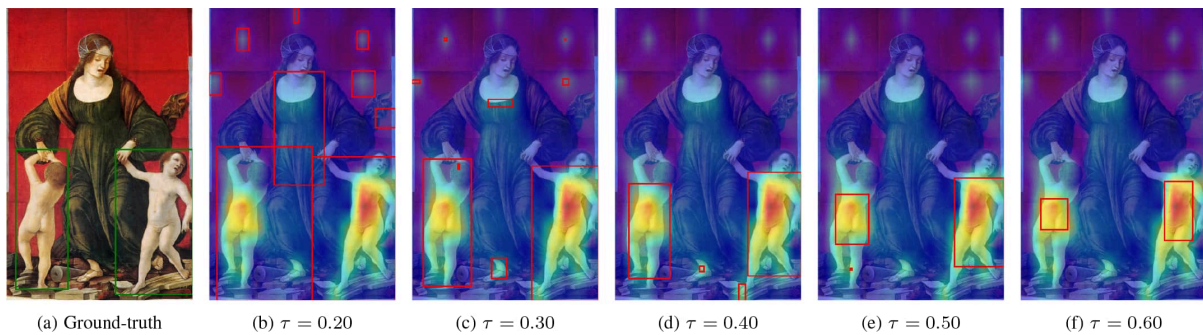


Figure 2: For Ercole de' Roberti's *The Wife of Hasdrubal and Her Children* (c. 1490 / 1493), the ground-truth bounding boxes for the concept «nudity» are shown in green (a). The bounding boxes derived from the class-conditional saliency map are displayed at progressively increasing threshold levels $\tau \in \{0.20, 0.30, 0.40, 0.50, 0.60\}$. [Visualization: Stefanie Schneider 2026]

³⁵ Cf. Choe et al. 2020.

³⁶ Cf. Choe et al. 2020.

All seven methods were evaluated using compatible CLIP variants: Grad-CAM, Grad-CAM++, LayerCAM, ScoreCAM, gScoreCAM, and CLIP Surgery were applied to the ResNet-50×16 backbone; LeGrad was evaluated on the ViT-B/32 backbone. For the gradient-based methods, features were extracted from the third ReLU activation in the final bottleneck block of the ResNet layer4, or from the last self-attention heads for *Vision Transformer (ViT)* models. Gradients of the class score with respect to these activations yielded the channel importance weights. For gradient-free methods, we masked each channel individually by setting all others in the upsampled activation map to zero and computed the cosine similarity between the resulting image embedding and the text prompt to derive the channel importance. CLIP Surgery applies two inference-time modifications – an adjusted self-attention block and a dual-path feed-forward network – to mitigate noisy visualizations without requiring any fine-tuning.³⁷ All code was implemented in PyTorch and ran on two NVIDIA GeForce RTX 2080 Ti.

[14]

3.2 Data

As Stefanie Schneider and Ricarda Vollmer have observed, object-level annotations – especially those specifying spatial localization through bounding boxes – are uncommon in art-historical datasets.³⁸ When such annotations do exist, they often deliberately avoid classes representing iconographic concepts, precisely those motifs whose semantic variability resists stable formalization. The DEArt dataset, for instance, is restricted to broad descriptive categories such as ›people‹ or to visually unambiguous but iconographically neutral objects like ›boats‹ or ›vases‹.³⁹ Yet precisely these unstable iconographic categories most urgently require closer examination if we are to assess whether VLMs can move beyond generic recognition to something resembling art-historical expertise.

[15]

For this reason, our evaluation focused on two datasets that explicitly provide iconographic content: IconArt⁴⁰ and ArtDL⁴¹ (Table 1; Fig. 3). Both provide annotations not only for figures (e.g., ›Saint Sebastian‹) but also for attributes and symbols (e.g., ›Ointment Jar‹). Of these, ArtDL is broader in scope, comprising 10 saints and 49 attributes. However, the usefulness of these datasets as benchmarks is limited by their distribution: Both have pronounced long-tail distributions, with a few frequent classes and many sparsely represented ones. In IconArt, three generic categories – ›beard‹ (21.98 %), ›angel‹ (21.15 %), and ›nudity‹ (15.39 %) – account for nearly two-thirds of all annotations, while classes of greater art-historical specificity such as ›Saint Sebastian‹ (1.66 %) or ›Crucifixion of Jesus‹ (2.21 %) appear only rarely. ArtDL displays the same dynamic even more strongly: The class ›face‹ (21.28 %) dominates more semantically charged labels, followed by ›Mary‹ (7.46 %) and ›Baby Jesus‹ (6.04 %). Attributes critical for saint identification – such as the ›lily‹ (0.92 %) – are relegated to statistical noise, with over half of the classes represented by fewer than 50 instances. The implications are methodological as much as statistical. Evaluating these corpora has the potential to reward models for recognizing abundant, generic classes while obscuring failures in categories of genuine iconographic expertise. In other words, a model may perform convincingly on aggregate measures while exhibiting no real grasp of the iconographic logics central to art-historical interpretation. However, both IconArt and ArtDL are indispensable precisely because they are the only datasets that even begin to capture iconographic detail. They are best understood, then, not as definitive benchmarks but as starting points – ›partial ground-truths‹, so to speak, that expose, as much as they enable, the epistemic challenges of employing VLMs in art history.

[16]

³⁷ Cf. Li et al. 2025.

³⁸ Cf. Schneider / Vollmer 2024.

³⁹ Cf. Reshetnikov et al. 2022.

⁴⁰ Cf. Gonthier et al. 2018.

⁴¹ Cf. Milani / Fraternali 2021.

Dataset	Images			Boxes	Classes
	Total	Positive	Negative		
ArtDL	4,166	808	3,358	3,793	59
IconArt	1,480	1037	443	4,931	10

Table 1: Statistics of the ArtDL (Milani / Fraternali 2021) and IconArt (Gonthier et al. 2018) test datasets. Positive images contain at least one ground-truth annotation; negative images contain none.

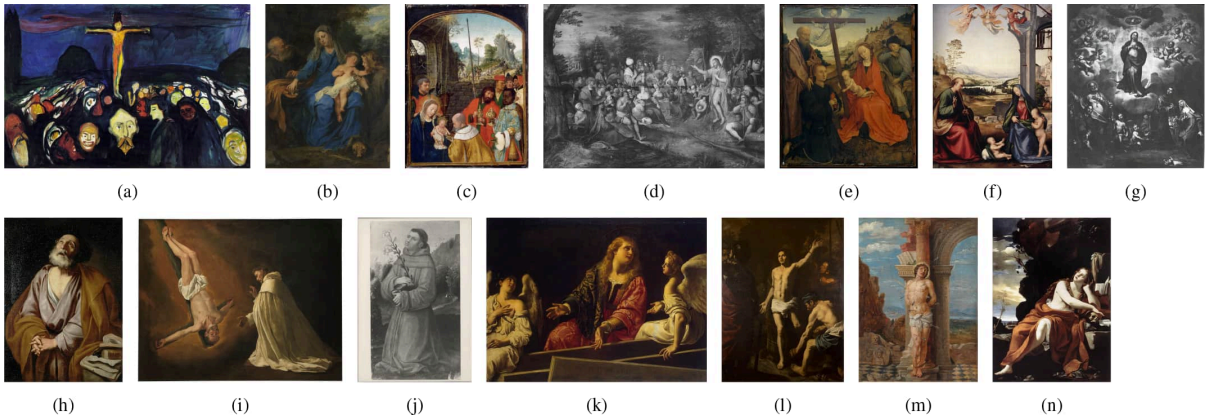


Figure 3: Selected images are shown from the IconArt (top row; Gonthier et al. 2018) and ArtDL (bottom row; Milani / Fraternali 2021) test sets. [Visualization: Stefanie Schneider 2026]

3.3 Results

Table 2 reports the comparative performance of all evaluated methods. Across both the IconArt and ArtDL test sets, CLIP Surgery achieved higher accuracy scores than all other methods, particularly at the more permissive IoU threshold of 0.30. On the ArtDL test set, it obtained a BoxAcc of 52.28 % at an IoU threshold of 0.30 – an absolute improvement of almost 9 points over the second-best method, LeGrad (43.82 %); this is also evident at the stricter IoU threshold of 0.50, where CLIP Surgery yielded a BoxAcc of 30.19 %. This advantage extended across object scales:⁴² At IoU ≥ 0.30, CLIP Surgery demonstrated the highest accuracy for small (20.69 %), medium (49.46 %), and large objects (74.87 %). We observed some exceptions at the class level – for instance, ›baby Jesus‹ performed better with LeGrad (76.79 %) than with CLIP Surgery (48.21 %) – but these are relatively isolated cases. Even when the threshold rose to IoU ≥ 0.50, where precise localization is more challenging, it remained the most accurate in all size categories. In contrast, gradient-based methods – Grad-CAM, Grad-CAM++, and LayerCAM – experienced significant performance degradation under both IoU thresholds. On the IconArt test set, CLIP Surgery similarly achieved the highest BoxAcc at IoU ≥ 0.30 (28.76 %); yet LeGrad slightly outperformed it in medium-object accuracy (25.85 % versus 24.47 %). In addition, object-level analysis revealed considerable differences between the two leading methods. For instance, in IconArt, CLIP Surgery recognized the class ›Mary‹ with an accuracy of 85.82 % for large objects, compared to 76.62 % with LeGrad. The difference is even greater for ›nudity‹, where CLIP Surgery attained 66.16 % versus LeGrad's 50.25 %. At the stricter IoU ≥ 0.50, CLIP Surgery reclaimed overall superiority (with a BoxAcc of 14.82 %), with LeGrad performing marginally better on small and medium objects. Gradient-based methods performed still more poorly here, underscoring their limited transferability to iconographic material.

⁴² Objects were grouped according to the area of their ground-truth bounding box relative to the total image area into small, medium, and large categories.

Dataset	Method	IoU ≥ 0.30				IoU ≥ 0.50			
		BoxAcc	BoxAcc _S	BoxAcc _M	BoxAcc _L	BoxAcc	BoxAcc _S	BoxAcc _M	BoxAcc _L
IconArt	CLIP Surgery	0.2876	0.0862	0.2447	0.6623	0.1482	0.0219	0.0807	0.4070
	LeGrad	0.2722	0.0849	0.2585	0.6436	0.1369	0.0232	0.0981	0.3808
	ScoreCAM	0.2411	0.0433	0.2264	0.6230	0.1040	0.0107	0.0852	0.2953
	gScoreCAM	0.2344	0.0679	0.2356	0.6024	0.1121	0.0174	0.0843	0.3208
	GradCAM	0.1391	0.0666	0.2145	0.2509	0.0355	0.0156	0.0660	0.0624
	GradCAM ++	0.1584	0.0277	0.0880	0.4432	0.0584	0.0067	0.0192	0.1779
	LayerCAM	0.1783	0.0420	0.1769	0.4694	0.0627	0.0094	0.0577	0.1823
ArtDL	CLIP Surgery	0.5228	0.2069	0.4946	0.7487	0.3019	0.0722	0.2205	0.5297
	LeGrad	0.4382	0.1788	0.4280	0.7035	0.2552	0.0458	0.1868	0.4680
	ScoreCAM	0.3557	0.0667	0.3055	0.6457	0.1672	0.0125	0.1087	0.3441
	gScoreCAM	0.3815	0.0889	0.3675	0.6627	0.1727	0.0208	0.1378	0.3497
	GradCAM	0.2684	0.1056	0.3851	0.2790	0.0701	0.0278	0.0988	0.0787
	GradCAM ++	0.2418	0.0403	0.1493	0.4890	0.1168	0.0069	0.0237	0.2484
	LayerCAM	0.2476	0.0583	0.2052	0.4578	0.0873	0.0097	0.0582	0.1590

Table 2: Bounding box detection results are reported for the IconArt (Gonthier et al. 2018) and ArtDL (Milani / Fraternali 2021) test sets, using the highest BoxAcc obtained across all binarization thresholds τ . The best performing approach per test set is indicated in bold. The subscripts S, M, and L denote ›small‹, ›medium‹, and ›large‹ objects, respectively.

The lower overall accuracy observed on IconArt relative to ArtDL arises from both structural and semantic differences between the two corpora. First, IconArt contains a higher proportion of small objects – those occupying $\leq 1\%$ of the image area – with 46.30 % of all instances, versus 18.98 % in ArtDL. Small objects are inherently more challenging to detect and classify, because they provide fewer pixels for feature extraction and are more susceptible to background clutter; the overrepresentation of small objects consequently reduces the aggregate performance on IconArt. Second, the datasets have different epistemic scopes. ArtDL is broader and more generic, including categories that can be recognized without specialized iconographic knowledge, e.g., ›beard‹. IconArt, by contrast, focuses on a small number of historically charged motifs – such as the ›Crucifixion of Jesus‹ – whose correct identification depends on contextual and narrative cues rather than isolated attributes. These scenes are formally and semantically dense: Their complexity and entanglement with related subthemes (e.g., episodes from Christ’s Passion) expose the limitations of models optimized for visual generality.

[18]

4. Case Study 2

So, what can we establish at this point? Across two large-scale datasets, our evaluation shows that CLIP Surgery consistently outperformed all other methods, with LeGrad emerging – almost unequivocally – as the second-best approach. Yet the discrepancies between the two datasets, IconArt and ArtDL, are instructive. While both annotate iconographic figures and attributes, they, inevitably, only partially reflect the broader iconographic heterogeneity that defines art-historical material. Our large-scale approach therefore provides valuable, albeit ultimately limited, perspectives as the selected classes naturally constrain the horizon of possible inquiry: We can assess performance within these constraints, but we cannot assume that the

[19]

epistemic space of art history is adequately represented.⁴³ The second case study thus deliberately shifted focus. It addressed the gap already identified by Miller, namely the absence of human-centered evaluations of XAI methods.⁴⁴ While the first case study focused on measuring the accuracy of these methods and assigning numerical performance scores, this second case study explored *interpretability* – asking not only *what* these models predict, but *how* their explanations were understood by human users. To approximate the practical variability of art-historical research – resistant as it is to fixed classification – we employed a broader and more diverse selection of images and categories. Our methodological scope remained consistent with the first study, encompassing the same explanation methods – Grad-CAM, Grad-CAM++, LayerCAM, LeGrad, ScoreCAM, gScoreCAM, and CLIP Surgery. But the focus was no longer on algorithmic performance. Rather, it tested whether these methods could make visible the complex visual logics that art history engages with, and whether they succeed, or fail, in approximating the visual concepts that constitute the field.

4.1 Experimental Design

To assess the degree to which algorithm-generated saliency maps align with human perceptions of visual importance, we conducted a within-subjects online study using [SoSci Survey](#) between June and July 2025.⁴⁵ After reviewing and accepting an informed consent form, participants provided socio-demographic information, including age, gender identity, education level, and professional status. On each subsequent page, they were shown one of seven artworks and asked to use their mouse to annotate regions they deemed relevant to a specified class (Fig. 4). These annotations served as the human reference (ground-truth) for the subsequent evaluation. Each artwork was paired with two target classes: Petrus Christus's *A Goldsmith in his Shop* (1449) with ›convex mirror‹ and ›girdle‹; Franz von Stuck's *Adam and Eve* (c. 1920) with ›arm outstretched‹ and ›snake‹; Antonello da Messina's *Calvary* (1475) with ›John‹ and ›thief‹; Claude Monet's *Japanese Footbridge* (1899) with ›bridge‹ and ›flower‹; Jean-Auguste-Dominique Ingres's *Oedipus and the Sphinx* (1808) with ›left foot‹ and ›Sphinx‹; Sandro Botticelli's *The Lamentation* (c. 1490) with ›sword‹ and ›Virgin Mary‹; Bartholomeus van der Helst's *The Musician* (1662) with ›lustful‹ and ›sheet music‹. The selection of artworks spanned a broad range of periods and styles, from Renaissance devotional works to early twentieth-century Symbolism. This diversity was intentional: It ensured that annotators were exposed to heterogeneous visual traditions and compositional strategies. Some target classes referred to discrete, visually localized elements (e.g., ›bridge‹), while others invoked symbolic or abstract categories (e.g., ›lustful‹). The participants then viewed seven saliency maps for each image-class pair, again generated by the following explanation methods: Grad-CAM, Grad-CAM++, LayerCAM, LeGrad, ScoreCAM, gScoreCAM, and CLIP Surgery. They were asked to order these maps according to how well they reflected the regions they had previously identified as important. This ranking served as a subjective measure of the alignment between human visual attention and the algorithmic saliency outputs. To minimize order effects, both the sequence of image-class pairs and the presentation order of saliency maps were independently randomized for each participant. Only those participants who completed at least 4 of the 14 ranking tasks were included in the final analysis.

[20]

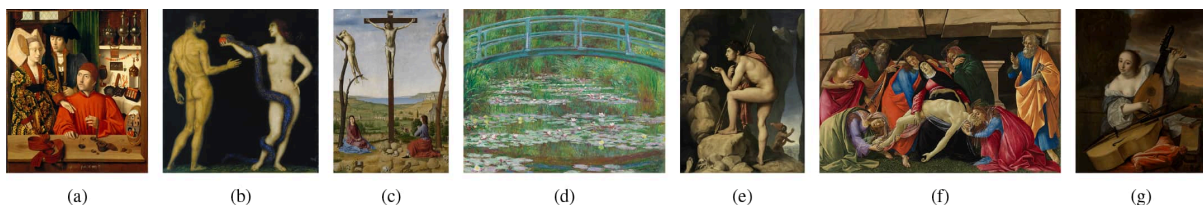


Figure 4: Artworks selected for the online study: Petrus Christus, *A Goldsmith in his Shop* (1449; a); Franz von Stuck, *Adam and Eve* (c. 1920; b); Antonello da Messina, *Calvary* (1475; c); Claude Monet, *Japanese Footbridge* (1899; d); Jean-Auguste-Dominique Ingres,

⁴³ This limitation, of course, applies to all machine-vision research, since available datasets – shaped by resource constraints – necessarily offer a highly selective view of historical material tailored to specific research questions.

⁴⁴ Cf. Miller 2019.

⁴⁵ The survey was last accessed on October 10, 2025.

Oedipus and the Sphinx (1808; e); Sandro Botticelli, *The Lamentation* (c. 1490; f); Bartholomeus van der Helst, *The Musician* (1662; g).
[Visualization: Stefanie Schneider 2026]

4.2 Participants

Participants were recruited through university-based channels at the University of Munich and the University of Göttingen; participation was voluntary and not incentivized. The study involved 33 participants, of whom 21.21 % identified as male, 75.76 % as female, and 3.03 % as diverse. Participant ages ranged from 18 to over 65 years (mean age = 42.21 years; standard deviation = 19.08 years). Regarding educational background, 39.39 % held qualifications for university entry, while 45.46 % had completed a degree at either a university or a university of applied sciences. Students made up the largest subgroup, accounting for 54.54 % of all respondents. In terms of art-historical expertise, 62.50 % reported basic knowledge (e.g., gained through introductory coursework), 21.88 % reported intermediate proficiency (gained through a bachelor's degree or initial professional experience), and the remainder described themselves as advanced or expert. The demographic profile thus reflects a fairly diverse sample in terms of both age and educational attainment. At the participant level, a complete-case analysis would have excluded 37.93 % of participants due to incomplete rankings for one or more image-class pairs; at the level of individual ranking tasks, 8.86 % were incomplete because not all seven methods had been ranked. We therefore imputed missing rank positions using the *Multivariate Imputation by Chained Equations (MICE)* algorithm over 20 iterations.⁴⁶ Although the sample size of $n = 33$ is at the lower end of recommended thresholds for within-subjects tests, power analysis indicates sufficient sensitivity to detect medium levels of inter-rater agreement (Kendall's coefficient of concordance $W \geq 0.30$) with an estimated probability of $\approx 80\text{--}85\%$. For the purposes of this investigation, such values are deemed adequate: The study was not intended to establish statistically significant differences between these specific artworks, but rather to assess – at least to some extent – the general ranking of saliency maps in terms of their alignment with human annotations. The central objective was to determine whether a modest, curated sample obtained through a user study could already reveal broader patterns that are informative for evaluating the applicability of XAI methods in art-historical contexts. [21]

4.3 Results

Fig. 5 shows divergent stacked bar charts of ranking distributions for each image–class pair. The charts illustrate the extent to which participants agreed or disagreed in their assessments, making it easier to identify patterns of consensus or divergence across images. To formally quantify inter-rater agreement, we report *Kendall's coefficient of concordance (W)* for each image–class pair: It ranges from 0 to 1, with higher values indicating stronger reliability and lower values greater variability in judgments. Across the images, three techniques – CLIP Surgery, LeGrad, and ScoreCAM – achieve the highest mean rank positions, indicating that participants perceived these methods as most faithfully highlighting their annotated regions; gScoreCAM also performs strongly, but shows slightly greater variability. In contrast, Grad-CAM, Grad-CAM++, and LayerCAM are consistently ranked towards the bottom, suggesting that their gradient-based heatmaps align less closely with the participants' own saliency judgments. [22]

⁴⁶ Cf. Buuren / Groothuis-Oudshoorn 2011.

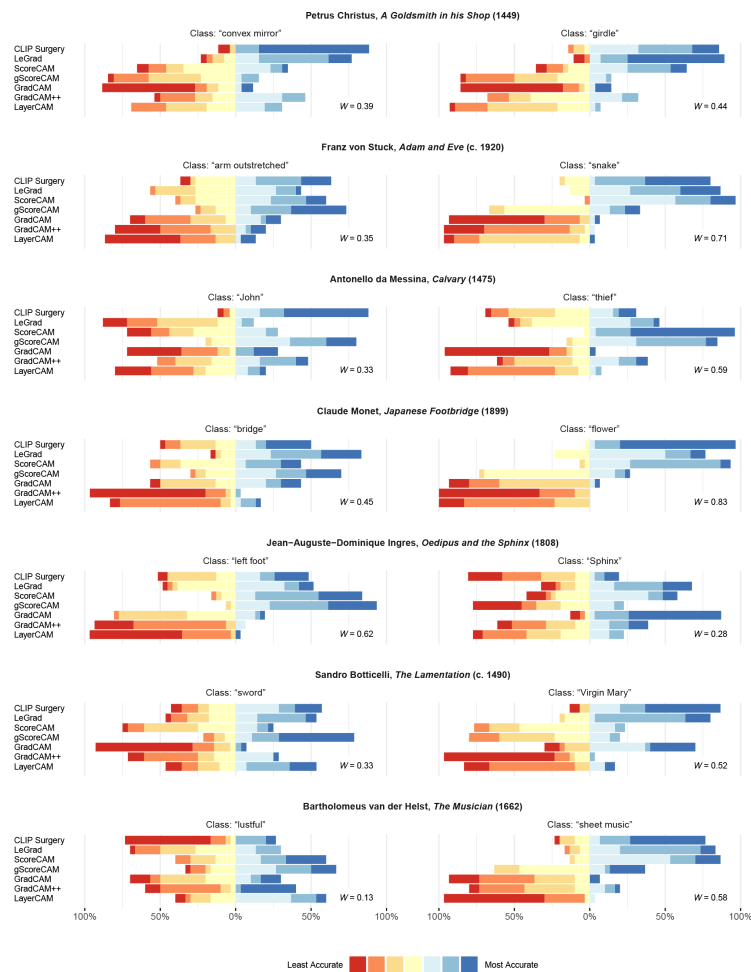


Figure 5: Evaluation results are shown as divergent stacked bar charts, comparing seven visual explainability methods for each image and class. Colors range from red (>least accurate<) to blue (>most accurate<). Kendall's W is reported to assess inter-rater reliability. [Visualization: Stefanie Schneider 2026]

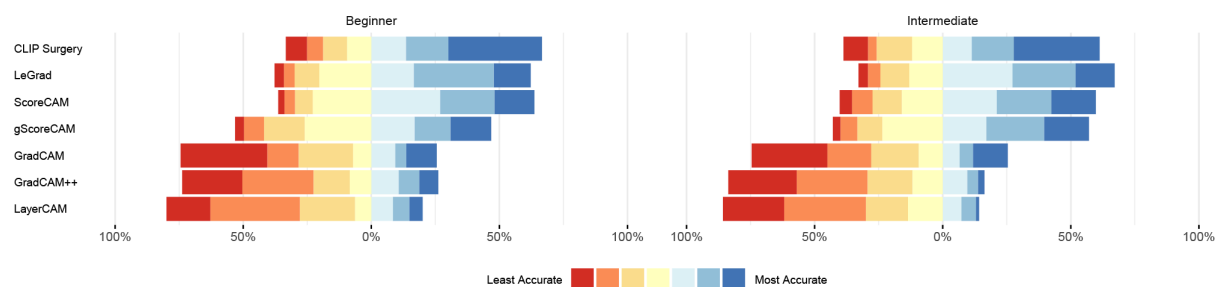


Figure 6: Evaluation results are shown as divergent stacked bar charts, comparing seven XAI methods across different levels of art-historical expertise. The results are aggregated over all images and classes. Colors range from red (>least accurate<) to blue (>most accurate<). [Visualization: Stefanie Schneider 2026]

When aggregated across all images and classes, these patterns were largely consistent across levels of art-historical expertise, as shown in Fig. 6. While CLIP Surgery was favored by participants with basic knowledge, those with at least intermediate proficiency had a slight preference for LeGrad; at this level, gScoreCAM also approached the top-performing group. Beyond these minor differences, no further – and in particular, no statistically significant – effects are observed. For visually well-defined or spatially localized targets – e.g., the

›snake‹ in Franz von Stuck's *Adam and Eve* (Fig. 7) or the ›left foot‹ in Jean-Auguste-Dominique Ingres's *Oedipus and the Sphinx* (Fig. 8) – participants' rankings converge tightly, with Kendall's W indicating strong inter-rater reliability ($W = 0.71$ and $W = 0.62$, respectively).

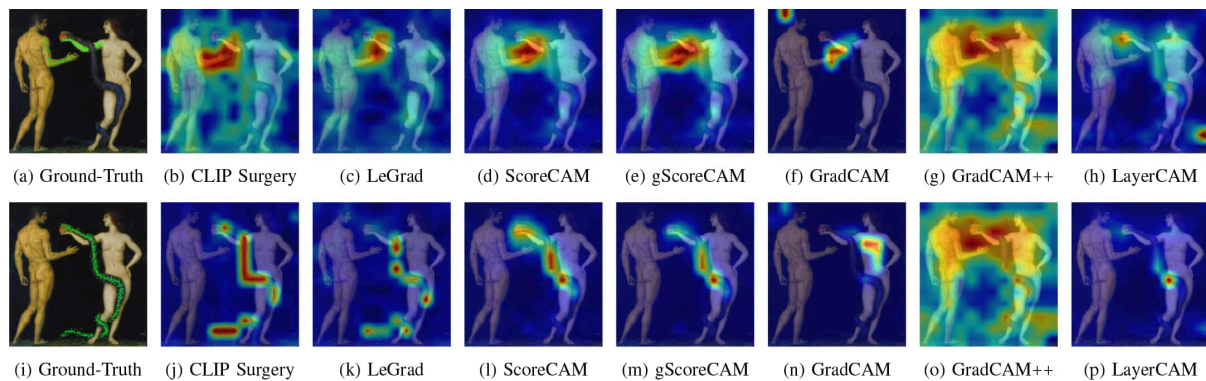


Figure 7: Ground-truth annotations and saliency maps for Franz von Stuck's *Adam and Eve* (c. 1920), shown for the classes ›arm outstretched‹ (top row) and ›snake‹ (bottom row). [Visualization: Stefanie Schneider 2026]

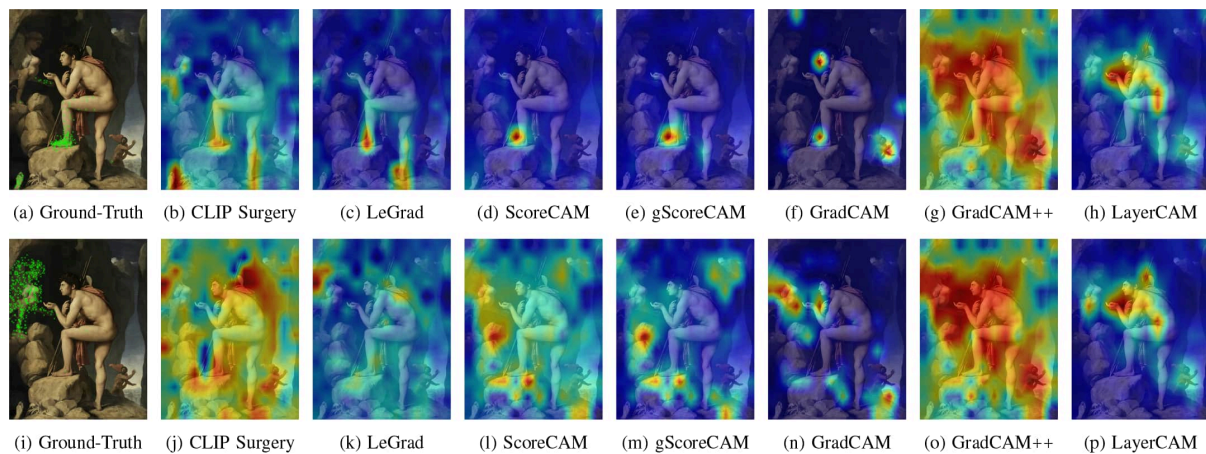


Figure 8: Ground-truth annotations and saliency maps for Jean-Auguste-Dominique Ingres's *Oedipus and the Sphinx* (1808), shown for the classes ›left foot‹ (top row) and ›Sphinx‹ (bottom row). [Visualization: Stefanie Schneider 2026]

A similar pattern emerges for the flowers beneath the bridge in Claude Monet's painting ($W = 0.83$), pointing to a shared perception of salient features (Fig. 9). By contrast, more diffuse or interpretative categories, such as ›lustful‹ (Fig. 10) or the ›Sphinx‹ (Fig. 8), yield widely dispersed rankings and low values Kendall's W, with no single method emerging as dominant; this reflects the intrinsic ambiguity of these higher-order visual concepts. However, the difficulty of some classes also becomes evident in cases where even human annotators struggle to establish consistent associations. Take, for example, Ingres's *Oedipus and the Sphinx*, in which three distinct ›left feet‹ would, in principle, need to be annotated: (1) The left foot of Oedipus, firmly bracing his weight; (2) the left foot of the Sphinx, obscured in shadow and therefore difficult to recognize – mirrored by the sparse human annotations in Fig. 8, and by the failure of saliency maps to capture it; and (3) a severed left foot at the lower edge of the canvas, possibly belonging to one of the Sphinx's earlier victims. In this case, annotators likely registered only the most prominent instance, suggesting insufficient time for close observation.

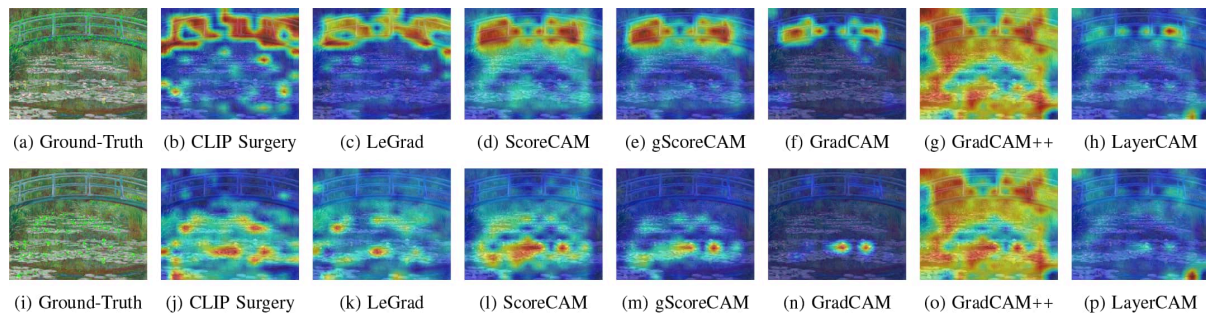


Figure 9: Ground-truth annotations and saliency maps for Claude Monet's *Japanese Footbridge* (1899), shown for the classes ›bridge‹ (top row) and ›flower‹ (bottom row). [Visualization: Stefanie Schneider 2026]

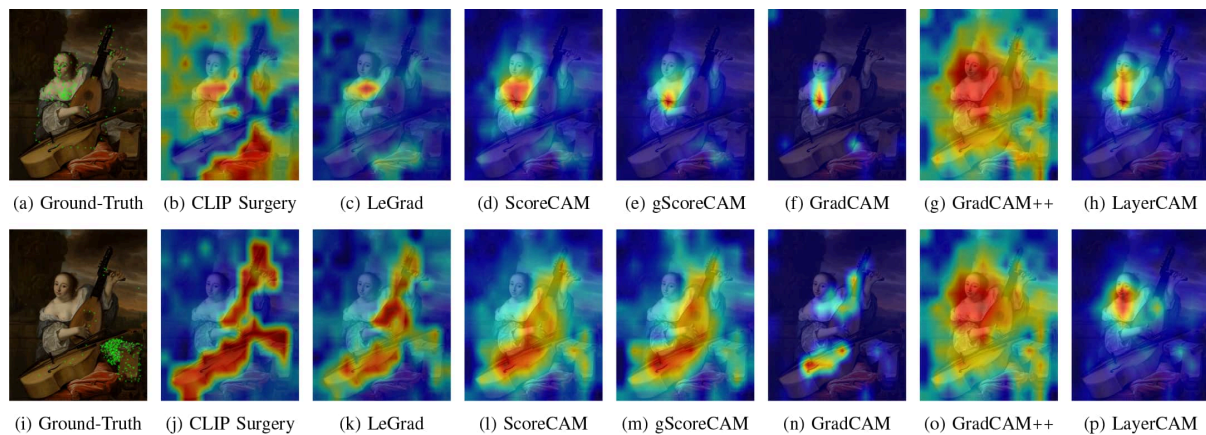


Figure 10: Ground-truth annotations and saliency maps for Bartholomeus van der Helst's *The Musician* (1662), shown for the classes ›lustful‹ (top row) and ›sheet music‹ (bottom row). [Visualization: Stefanie Schneider 2026]

Yet in other examples, the challenge arose less from attention than from limited knowledge of the relevant visual structures associated with a given class. An interesting case is the red wedding girdle in Petrus Christus's *Goldsmith*, which projects into the viewer's space over the shop's ledge, yet is far less frequently annotated than the woman's hip belt (Fig. 11). The issue is even more evident in Sandro Botticelli's *Lamentation*. While the dead Christ rests in the Virgin Mary's lap, two additional figures – Mary Magdalene and, very likely, Mary of Clopas – gently support his head and feet. Yet participants frequently mislabeled these figures as the Virgin Mary as well (Fig. 12).

[25]

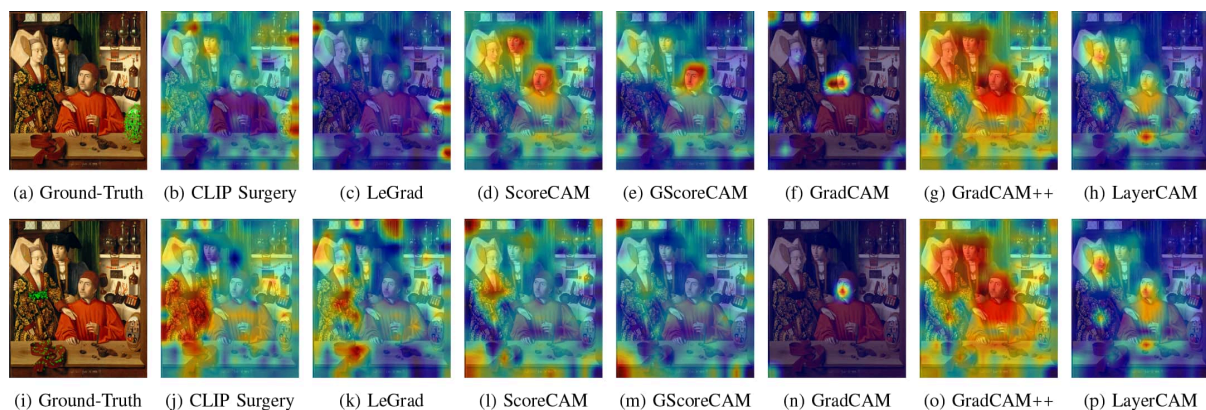


Figure 11: Ground-truth annotations and saliency maps for Petrus Christus's *A Goldsmith in his Shop* (1449), shown for the classes ›convex mirror‹ (top row) and ›girdle‹ (bottom row). [Visualization: Stefanie Schneider 2026]

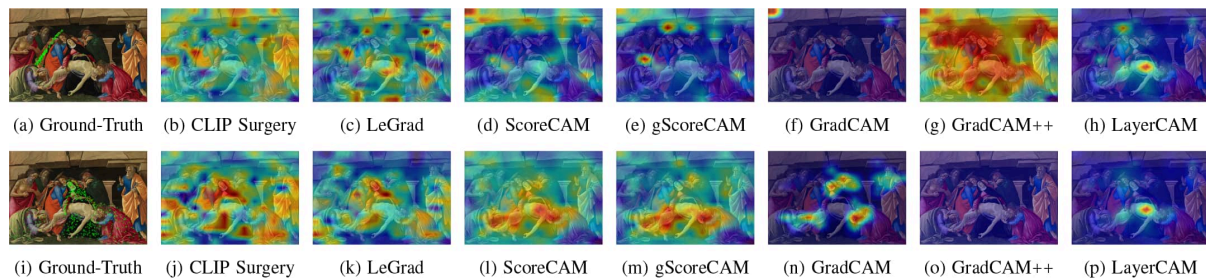


Figure 12: Ground-truth annotations and saliency maps for Sandro Botticelli's *The Lamentation* (c. 1490), shown for the classes ›sword‹ (top row) and ›Virgin Mary‹ (bottom row). [Visualization: Stefanie Schneider 2026]

5. Discussion

The case studies reveal two interrelated factors. (1) *Ambiguity of concepts*: In art-historical imagery, target classes often are interpretive rather than fixed, and thus resist stable localization. In Botticelli's *Lamentation*, the three Marys mourning over Christ are visually similar, so that non-specialist annotators might confuse them (Fig. 12). Here, *discernibility* emerges as a criterion: Where classes are both concrete and spatially bounded, annotators converge; where they are symbolic, context-dependent, or reliant on art-historical expertise, judgments diverge and inter-rater reliability declines. Ground-truth annotations, in this sense, are never exhaustive. Methods that highlight only the most visually dominant instance of a class therefore cannot easily be dismissed as inadequate, since they nonetheless establish a valid link between text and image. Yet ambiguity does not arise only from annotators: It is also encoded in the model itself. [26]

This leads to: (2) *Limits of representation*: Saliency methods cannot recover what the model itself fails to encode. If a concept does not appear as a localized hotspot in CLIP's latent space, attribution will necessarily remain vague, regardless of the post-hoc technique employed. This is evident in Antonello da Messina's *Calvary*: The thieves flanking Christ are inconsistently mapped – ScoreCAM and gScoreCAM weakly associate both figures with the prompt ›thief‹, while LeGrad isolates only the feet of the right-hand figure (Fig. 13). Such incoherence suggests that CLIP does not encode ›thief‹ as a transferable visual concept. The difficulty may partly stem from the unnatural posture of the figures, but – more fundamentally – it reflects the fragmentary nature of the training corpus. In contrast to the photographic imagery that dominates CLIP's dataset, crucifixion scenes are relatively uncommon, with peripheral figures such as the thieves being especially under-specified. The challenge is further complicated by the semantic breadth of the term ›thief‹. Unlike ›Christ‹ or ›cross‹, which correspond to highly codified and visually stable iconographic forms, ›thief‹ has no fixed template: It may denote an anonymous criminal, a masked burglar, or – in the Passion narrative – the unnamed ›good‹ and ›bad‹ thieves, who are distinguished only by their position relative to Christ and, in some traditions, by subtle gestures or expressions. Ambiguity, in this case, is not simply a problem of human perception, but a structural property of the machine's representational logic. Yet across both cases, relative performance of gradient-based, score-based, and hybrid methods remained largely stable. [27]

This indicates that small-scale, human-centered studies can support robust comparative evaluation, while large-scale evaluation on pre-existing datasets can extend such results – though only for a narrow subset of art-historically relevant categories. More importantly, these studies also function diagnostically: They reveal not only how well attribution methods capture model activations, but also how models themselves mediate art-historical concepts. Human studies, carefully designed, thus can produce findings that generalize to broader validations while simultaneously exposing the cultural and epistemic imaginaries embedded in machine vision. [28]

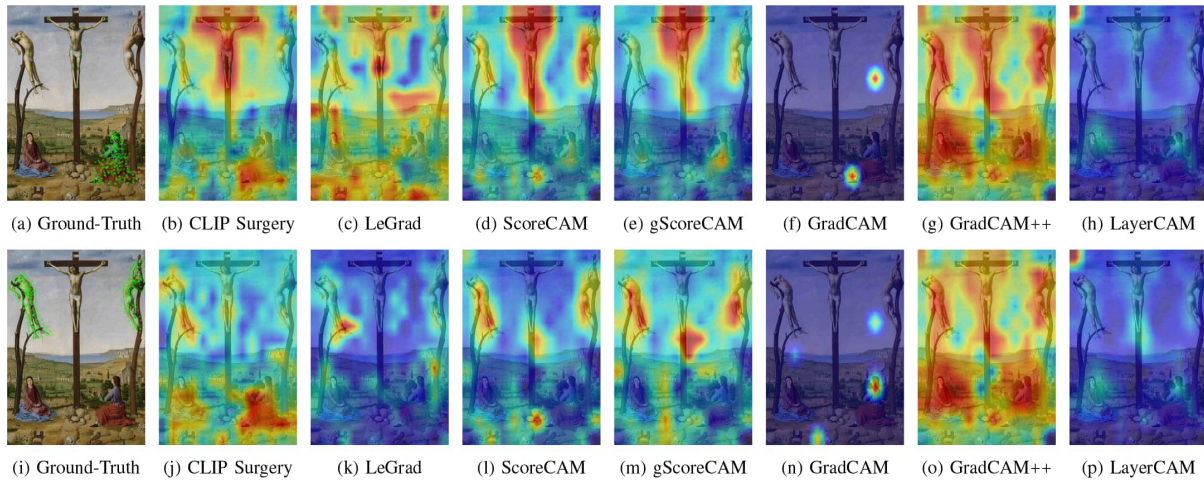


Figure 13: Ground-truth annotations and saliency maps for Antonello da Messina's *Calvary* (1475), shown for the classes ›John‹ (top row) and ›thief‹ (bottom row). [Visualization: Stefanie Schneider 2026]

For real-time or ad hoc explanation needs, additional considerations are required. ScoreCAM obtains channel weights by computing the forward-pass score for each activation map – i.e., it performs one forward pass per channel (i.e., C forward passes per image) – to generate a single heatmap (e.g., 3,072 forward passes for RN50×16), which makes it slow for on-the-fly explanations.⁴⁷ gScoreCAM alleviates this bottleneck by selecting only the top- k channels (typically $k = 300$), reducing the number of forward passes to about $0.1C \approx 307$ for $C = 3,072$.⁴⁸ By contrast, gradient-based methods (Grad-CAM, Grad-CAM++, LayerCAM, and variants such as LeGrad) require one forward pass and one backward pass to compute CAMs. However, for real-time or ad hoc explanations, CLIP Surgery is particularly advantageous, as it requires only a single modified forward pass without any gradient computations or backpropagation.⁴⁹ Beyond computational efficiency, these methodological differences also influence the epistemic character of the resulting explanations. Multi-pass methods like ScoreCAM and gScoreCAM often produce smoother, less noisy saliency maps, but their cost makes them impractical for interactive settings. Gradient-based methods, while faster, can suffer from gradient saturation or overemphasis on high-level features, resulting in instability or sensitivity across inputs. CLIP Surgery, though extremely efficient, forgoes gradient information entirely; its outputs should thus be interpreted as approximations of model ›attention‹ rather than as exhaustive mappings of activation contributions. In art-historical contexts, this distinction is nontrivial: When interpreting ambiguous or symbolical imagery, the choice of method not only constrains computational performance but also shapes the interpretive claims about what the model *perceives*. Real-time explanation pipelines, then, must balance latency constraints with the need for stable, interpretable visualizations – accepting that faster methods may favor responsiveness and accessibility, while slower methods may deliver higher resolution or fidelity but at the cost of usability in exploratory settings.

6. Conclusion

Our case studies – addressing the questions raised in Section 1 – demonstrated that the epistemic promise of XAI in digital art history is both methodological and hermeneutic. With regard to the first question (*How effectively do XAI methods localize iconographic objects in artworks under zero-shot conditions without fine-tuning?*), our quantitative evaluation has shown that CLIP Surgery, which is specifically adapted to CLIP's dual-encoder architecture, consistently outperformed general-purpose gradient- or score-based methods.

⁴⁷ Cf. Wang et al. 2020.

⁴⁸ Cf. Chen et al. 2022.

⁴⁹ Cf. Li et al. 2025. Actual runtime performance depends on hardware and implementation details. We recommend empirical benchmarking of all methods on the target system to obtain accurate latency estimates before deployment.

Even when trained primarily on non-art-historical imagery, CLIP Surgery could delineate a broad range of iconographic concepts, particularly when the target classes are visually distinct or spatially constrained; its accuracy, however, declined for semantically complex motifs. The limitation here lies not in the XAI method itself but in the representational granularity of the embedding space: CLIP does not, in fact, *see* an object in its historicity, but only the statistical residue of an already mediated image-world. Turning to the second question (*Does the visual relevance of these maps correspond to human judgments?*), our user study partially confirmed this observation. Participants generally preferred CLIP Surgery, LeGrad, and ScoreCAM, whose saliency maps closely approximated human annotations. Yet for abstract or context-dependent classes, agreement and, thus, inter-rater reliability decreased. This divergence underscores a central tension: While saliency maps can reproduce certain aspects of perceptual attention, they cannot replicate the interpretive depth of the art-historical gaze. The third question (*Which factors, such as object size and concept abstraction, drive performance differences?*) reveals both structural and semantic determinants. Localization accuracy correlates strongly with object size – and, thus, visual prominence – but it is equally influenced by conceptual stability. Classes with more consistent visual referents (e.g., ›bridge‹ or ›snake‹) yielded higher accuracy scores than those with vague, symbolic, or art-historically specific meanings (e.g., ›lustful‹ or ›Virgin Mary‹).

These findings suggest that model performance reflects how, and to what extent, a concept exists within the model's learned visual ontology. Returning to a broader question raised in [Section 1](#) – whether XAI discloses a model's internal conceptual structure or merely aestheticizes its opacity – our answer must be: It depends. The visual legibility of a saliency map can be deceptive as it does not imply epistemic transparency. Such maps expose the internal dynamics of how certain features or tokens activate within an embedding space, while concealing the historical, cultural, and linguistic priors that render those activations meaningful. What is visualized, therefore, is not the model's *understanding* of an artwork but the projection of human interpretive desire onto computational artifacts that can only approximate meaning. Explainability in digital art history must, in this context, be conceived as a dialogical process between human and machine vision – as elaborated by Miller⁵⁰ – demanding critical awareness of the epistemic imaginaries through which models *see*. XAI outputs should therefore be read not as self-sufficient explanations but as prompts for further hermeneutic inquiry.

[31]

Acknowledgements

This work was funded in part by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG) under project number [510048106](#). I thank Hubertus Kohle, Ralph Ewerth, Eric Müller-Budack, Matthias Springstein, and Julian Stalter for their insightful discussions and helpful comments on the subject matter.

[32]

⁵⁰ Cf. Miller 2019.

Bibliography

- Samira Abnar / Willem H. Zuidema: Quantifying Attention Flow in Transformers. In: Dan Jurafsky / Joyce Chai / Natalie Schluter / Joel R. Tetreault (eds.): Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online, 05.–10.07.2020). Stroudsburg, US-PA 2020, pp. 4190–4197. DOI: [10.18653/v1/2020.acl-main.385](https://doi.org/10.18653/v1/2020.acl-main.385)
- Emily M. Bender / Timnit Gebru / Angelina McMillan-Major / Shmargaret Shmitchell: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Madeleine Clare Elish / William Isaac / Richard S. Zemel (eds.): FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Canada, 03.03.–10.03.2021). New York 2021, pp. 610–623. DOI: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922)
- Usha Bhalla / Alex Oesterling / Suraj Srinivas / Flávio P. Calmon / Himabindu Lakkaraju: Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE). In: Amir Globerson / Lester Mackey / Danielle Belgrave / Angela Fan / Ulrich Paquet / Jakub M. Tomczak / Cheng Zhang (eds.): Advances in Neural Information Processing Systems 37. NeurIPS 2024. 38th Annual Conference on Neural Information Processing Systems (Vancouver, 10.12.–15.12.2024). 2024. [\[online\]](#)
- Abeba Birhane / Vinay Uday Prabhu / Emmanuel Kahembwe: Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes. arXiv. 05.10.2021. DOI: [10.48550/arXiv.2110.01963](https://doi.org/10.48550/arXiv.2110.01963)
- Walid Bousselham / Angie W. Boggust / Sofian Chayboudi / Hendrik Strobelt / Hilde Kuehne: LeGrad: An Explainability Method for Vision Transformers via Feature Formation Sensitivity. arXiv. 04.04.2024. Version 2: 08.01.2025. DOI: [10.48550/arXiv.2404.03214v2](https://doi.org/10.48550/arXiv.2404.03214v2)
- Bruce G. Buchanan / Edward H. Shortliffe (eds.): Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, US-MA etc. 1985. [\[GVK record\]](#)
- Stef van Buuren / Karin Groothuis-Oudshoorn: mice: Multivariate Imputation by Chained Equations in R. In: Journal of Statistical Software 45 (2011), no. 3, pp. 1–67. DOI: [10.18637/jss.v045.i03](https://doi.org/10.18637/jss.v045.i03)
- Aditya Chattopadhyay / Anirban Sarkar / Prantik Howlader / Vineeth N. Balasubramanian: Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV; Lake Tahoe, US-NV, 12.03.–15.03.2018). 2018, pp. 839–847. DOI: [10.1109/WACV.2018.00097](https://doi.org/10.1109/WACV.2018.00097)
- Peijie Chen / Qi Li / Saad Biaz / Trung Bui / Anh Nguyen: gScoreCAM: What Objects Is CLIP Looking At? In: Lei Wang / Juergen Gall / Tat-Jun Chin / Imari Sato / Rama Chellappa (eds.): Computer Vision – ACCV 2022: 16th Asian Conference on Computer Vision (= Lecture Notes in Computer Science, 13844; Macao, 04.12.–08.12.2022). Cham 2022, pp. 588–604. DOI: [10.1007/978-3-031-26316-3_35](https://doi.org/10.1007/978-3-031-26316-3_35)
- Junsuk Choe / Seong Joon Oh / Seungho Lee / Sanghyuk Chun / Zeynep Akata / Hyunjung Shim: Evaluating Weakly Supervised Object Localization Methods Right. In: 2020 IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020; Seattle, 13.06.–19.06.2020). New York 2020, pp. 3130–3139. DOI: [10.1109/CVPR42600.2020.00320](https://doi.org/10.1109/CVPR42600.2020.00320)
- Wolfgang Coy / Lena Bonsiepen: Erfahrung und Berechnung: Kritik der Expertensystemtechnik. Berlin etc. 1989. [\[GVK record\]](#)
- Nicolas Gonthier / Yann Gousseau / Said Ladjal / Olivier Bonfait: Weakly Supervised Object Detection in Artworks. In: Laura Leal-Taixé / Stefan Roth (eds.): Computer Vision – ECCV 2018 Workshops (= Lecture Notes in Computer Science, 11130; Munich, 08.09.–14.09.2018). Cham 2018, pp. 692–709. DOI: [10.1007/978-3-030-11012-3_53](https://doi.org/10.1007/978-3-030-11012-3_53)
- Tanmay Gupta / Arash Vahdat / Gal Chechik / Xiaodong Yang / Jan Kautz / Derek Hoiem: Contrastive Learning for Weakly Supervised Phrase Grounding. In: Andrea Vedaldi / Horst Bischof / Thomas Brox / Jan-Michael Frahm (eds.): Computer Vision – ECCV 2020: 16th European Conference (= Lecture Notes in Computer Science, 12348; Glasgow, 23.08.–28.08.2020). Cham 2020, pp. 752–768. DOI: [10.1007/978-3-030-58580-8_44](https://doi.org/10.1007/978-3-030-58580-8_44)
- Paul Harmon / Rex Maus / William Morrissey: Expertensysteme: Werkzeuge und Anwendungen. München etc. 1989. [\[GVK record\]](#)
- Vikas Hassija / Vinay Chamola / Atmesh Mahapatra / Abhinandan Singal / Divyansh Goel / Kaizhu Huang / Simone Scardapane / Indro Spinelli / Mufti Mahmud / Amir Hussain: Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. In: Cognitive Computation 16 (2024), no. 1, pp. 45–74. DOI: [10.1007/s12559-023-10179-8](https://doi.org/10.1007/s12559-023-10179-8)
- Leonardo Impett / Fabian Offert: There Is a Digital Art History. In: Visual Resources 38 (2022), no. 2, pp. 186–209. DOI: [10.1080/01973762.2024.2362466](https://doi.org/10.1080/01973762.2024.2362466)
- Sarthak Jain / Byron C. Wallace: Attention is not Explanation. In: Jill Burstein / Christy Doran / Thamar Solorio (eds.): Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). Minneapolis 2019, pp. 3543–3556. DOI: [10.18653/v1/N19-1357](https://doi.org/10.18653/v1/N19-1357)
- Peng-Tao Jiang / Chang-Bin Zhang / Qibin Hou / Ming-Ming Cheng / Yunchao Wei: LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. In: IEEE Transactions on Image Processing 30 (2021), pp. 5875–5888. DOI: [10.1109/TIP.2021.3089943](https://doi.org/10.1109/TIP.2021.3089943)
- Rémi Kazmierczak / Eloise Berthier / Goran Frehse / Gianni Franchi: CLIP-QDA: An Explainable Concept Bottleneck Model. In: Transactions on Machine Learning Research 05 (2024). arXiv. 30.11.2023. Version 3: 31.05.2024. DOI: [abs/2312.00110v3](https://doi.org/abs/2312.00110v3)
- Junnan Li / Dongxu Li / Silvio Savarese / Steven C. H. Hoi: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: Andreas Krause / Emma Brunskill / Kyunghyun Cho / Barbara Engelhardt / Sivan Sabato / Jonathan Scarlett (eds.): Proceedings of the 40th International Conference on Machine Learning (Honolulu, US-HI, 23.07.–29.07.2023). PMLR 202 (2023), pp. 19730–19742. [\[online\]](#)
- Yi Li / Hualiang Wang / Yiqun Duan / Jiheng Zhang / Xiaomeng Li: A Closer Look at the Explainability of Contrastive Language-Image Pre-Training. In: Pattern Recognition 162 (2025). DOI: [10.1016/j.patcog.2025.111409](https://doi.org/10.1016/j.patcog.2025.111409)
- Federico Milani / Piero Fraternali: A Dataset and a Convolutional Model for Iconography Classification in Paintings. In: ACM Journal on Computing and Cultural Heritage 14 (2021), no. 4, pp. 1–18. DOI: [10.1145/3458885](https://doi.org/10.1145/3458885)
- Tim Miller: Explanation in Artificial Intelligence: Insights from the Social Sciences. In: Artificial Intelligence 267 (2019), pp. 1–38. DOI: [10.1016/j.artint.2018.07.007](https://doi.org/10.1016/j.artint.2018.07.007)
- Fabian Offert / Peter Bell: imgs.ai: A Deep Visual Search Engine for Digital Art History. In: Anne Bailiot / Toma Tasovac / Walter Scholger / Georg Vogeler (eds.): DH 2023. Collaboration as Opportunity. International Conference of the Alliance of Digital Humanities Organizations (Graz, 10.07.–14.07.2023). Graz 2023. DOI: [10.5281/zenodo.8107778](https://doi.org/10.5281/zenodo.8107778)
- Tuomas Oikarinen / Tsui-Wei Weng: CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks. In: The 11th International Conference on Learning Representations (Kigali, Rwanda, 01.05.–05.05.2023). 2023. [\[online\]](#)
- Frank Puppe: Einführung in Expertensysteme. Berlin etc. 1988. [\[GVK record\]](#)
- Alec Radford / Jong Wook Kim / Chris Hallacy / Aditya Ramesh / Gabriel Goh / Sandhini Agarwal / Girish Sastry / Amanda Askell / Pamela Mishkin / Jack Clark / Gretchen Krueger / Ilya Sutskever: Learning Transferable Visual Models From Natural Language Supervision. In: Marina Meila / Tong Zhang (eds.): Proceedings of the 38th International Conference on Machine Learning (Virtual, 18.07.–24.07.2021). PMLR 139 (2021), pp. 8748–8763. [\[online\]](#)

- Artem Reshetnikov / Maria-Cristina V. Marinescu / Joaquim Moré López: DEArt: Dataset of European Art. In: Leonid Karlinsky / Tomer Michaeli / Ko Nishino (eds.): Computer Vision – ECCV 2022 Workshops (= Lecture Notes in Computer Science, 13801; Tel Aviv, Israel, 23.10.–27.10.2022). Cham 2022, pp. 218–233. DOI: 10.1007/978-3-031-25056-9_15
- Waddah Saeed / Christian W. Omlin: Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities. In: Knowledge-Based Systems 263 (2023). DOI: 10.1016/j.knosys.2023.110273
- Stefanie Schneider / Ricarda Vollmer: Poses of People in Art: A Dataset for Human Pose Estimation in Digital Art History. In: ACM Journal on Computing and Cultural Heritage 17 (2024), no. 4, pp. 1–19. DOI: [10.1145/3696455](https://doi.org/10.1145/3696455)
- Christoph Schuhmann / Richard Vencu / Romain Beaumont / Robert Kaczmarczyk / Clayton Mullis / Aarush Katta / Theo Coombes / Jenia Jitsev / Aran Komatsuzaki: LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv. 03.11.2021. DOI: [10.48550/arXiv.2111.02114](https://doi.org/10.48550/arXiv.2111.02114)
- Ramprasaath R. Selvaraju / Michael Cogswell / Abhishek Das / Ramakrishna Vedantam / Devi Parikh / Dhruv Batra: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV 2017; Venice, 22.10.–29.10.2017). New York 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74
- Timo Speith: A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In: FAcT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (Seoul, 21.06.–24.06.2022). New York 2022, pp. 2239–2250. DOI: 10.1145/3531146.3534639
- Matthias Springstein / Stefanie Schneider / Javad Rahnama / Eyke Hüllermeier / Hubertus Kohle / Ralph Ewerth: iART: A Search Engine for Art-Historical Images to Support Research in the Humanities. In: Heng Tao Shen / Yueting Zhuang / John R. Smith / Yang Yang / Pablo César / Florian Metze / Balakrishnan Prabhakaran (eds.): MM '21: Proceedings of the 29th ACM International Conference on Multimedia (Virtual, 20.10.–24.10.2021). New York 2021, pp. 2801–2803. DOI: 10.1145/3474085.3478564
- Haofan Wang / Zifan Wang / Mengnan Du / Fan Yang / Zijian Zhang / Sirui Ding / Piotr Mardziel / Xia Hu: Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In: 2020 IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR Workshops 2020). Seattle 2020, pp. 111–119. DOI: 10.1109/CVPRW50498.2020.00020
- Xiaohua Zhai / Xiao Wang / Basil Mustafa / Andreas Steiner / Daniel Keysers / Alexander Kolesnikov / Lucas Beyer: LiT: Zero-Shot Transfer with Locked-Image Text Tuning. In: IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022). New Orleans 2022, pp. 18102–18112. DOI: 10.1109/CVPR52688.2022.01759

List of Figures and Tables

- Figure 1: The saliency map highlights, in red, the image regions most strongly associated with the concept of the ›snake‹ in Franz von Stuck's *Adam and Eve* (c. 1920). [Visualization: Stefanie Schneider 2026]
- Figure 2: For Ercole de' Roberti's *The Wife of Hasdrubal and Her Children* (c. 1490 / 1493), the ground-truth bounding boxes for the concept ›nudity‹ are shown in green (a). The bounding boxes derived from the class-conditional saliency map are displayed at progressively increasing threshold levels $\tau \in \{0.20, 0.30, 0.40, 0.50, 0.60\}$. [Visualization: Stefanie Schneider 2026]
- Table 1: Statistics of the ArtDL (Milani / Fraternali 2021) and IconArt (Gonthier et al. 2018) test datasets. Positive images contain at least one ground-truth annotation; negative images contain none.
- Figure 3: Selected images are shown from the IconArt (top row; Gonthier et al. 2018) and ArtDL (bottom row; Milani / Fraternali 2021) test sets. [Visualization: Stefanie Schneider 2026]
- Table 2: Bounding box detection results are reported for the IconArt (Gonthier et al. 2018) and ArtDL (Milani / Fraternali 2021) test sets, using the highest BoxAcc obtained across all binarization thresholds τ . The best performing approach per test set is indicated in bold. The subscripts S, M, and L denote ›small‹, ›medium‹, and ›large‹ objects, respectively.
- Figure 4: Artworks selected for the online study: Petrus Christus, *A Goldsmith in his Shop* (1449; a); Franz von Stuck, *Adam and Eve* (c. 1920; b); Antonello da Messina, *Calvary* (1475; c); Claude Monet, *Japanese Footbridge* (1899; d); Jean-Auguste-Dominique Ingres, *Oedipus and the Sphinx* (1808; e); Sandro Botticelli, *The Lamentation* (c. 1490; f); Bartholomeus van der Helst, *The Musician* (1662; g). [Visualization: Stefanie Schneider 2026]
- Figure 5: Evaluation results are shown as divergent stacked bar charts, comparing seven visual explainability methods for each image and class. Colors range from red (least accurate) to blue (most accurate). Kendall's W is reported to assess inter-rater reliability. [Visualization: Stefanie Schneider 2026]
- Figure 6: Evaluation results are shown as divergent stacked bar charts, comparing seven XAI methods across different levels of art-historical expertise. The results are aggregated over all images and classes. Colors range from red (least accurate) to blue (most accurate). [Visualization: Stefanie Schneider 2026]
- Figure 7: Ground-truth annotations and saliency maps for Franz von Stuck's *Adam and Eve* (c. 1920), shown for the classes ›arm outstretched‹ (top row) and ›snake‹ (bottom row). [Visualization: Stefanie Schneider 2026]
- Figure 8: Ground-truth annotations and saliency maps for Jean-Auguste-Dominique Ingres's *Oedipus and the Sphinx* (1808), shown for the classes ›left foot‹ (top row) and ›Sphinx‹ (bottom row). [Visualization: Stefanie Schneider 2026]
- Figure 9: Ground-truth annotations and saliency maps for Claude Monet's *Japanese Footbridge* (1899), shown for the classes ›bridge‹ (top row) and ›flower‹ (bottom row). [Visualization: Stefanie Schneider 2026]
- Figure 10: *The Musician* (1662), shown for the classes ›lustful‹ (top row) and ›sheet music‹ (bottom row). [Visualization: Stefanie Schneider 2026]
- Figure 11: *A Goldsmith in his Shop* (1449), shown for the classes ›convex mirror‹ (top row) and ›girdle‹ (bottom row). [Visualization: Stefanie Schneider 2026]
- Figure 12: *The Lamentation* (c. 1490), shown for the classes ›sword‹ (top row) and ›Virgin Mary‹ (bottom row). [Visualization: Stefanie Schneider 2026]
- Figure 13: Ground-truth annotations and saliency maps for Antonello da Messina's *Calvary* (1475), shown for the classes ›john‹ (top row) and ›thief‹ (bottom row). [Visualization: Stefanie Schneider 2026]