Understanding How MLLMs Describe Artworks Using Token Activation Maps

Nicola Fanelli[1 [0009−0007−6602−7504]], Pasquale De Marinis[1 [0000−0001−8935−9156]], Raffaele Scaringi[1 [0000−0001−7512−7661]], Eva Cetinic[2 [0000−0002−5330−1259]], Gennaro Vessio[1 [0000−0002−0883−2691]], and Giovanna Castellano[1 [0000−0002−6489−8628]]

[1] Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
{nicola.fanelli, gennaro.vessio, pasquale.demarinis, raffaele.scaringi, giovanna.castellano}@uniba.it
[2] Digital Society Initiative, University of Zurich, Zurich, Switzerland
eva.cetinic@uzh.ch

Fig. 1: **We present a token-level view of how multimodal LLMs *see* the art they describe.** Using Token Activation Maps, we trace each generated span back to the image region it draws on. Grounding depends on *what* is said: a concrete subject (CVO, *"small village or town"*) localizes to one region, while a style descriptor (STYLE, *"dynamic brushstrokes"*) and a metadata mention (META, *"Starry Night"*) spread diffusely across the canvas.

Abstract. Multimodal Large Language Models (MLLMs) describe artworks with remarkable fluency, yet the visual reasoning behind their outputs remains opaque. When an MLLM names a style, identifies a subject, or recognizes an iconographic symbol, does it ground each claim in the relevant region of the canvas, draw on an undifferentiated visual signal, or rely primarily on textual priors? We study this using the Token Activation Map (TAM), which produces, for each generated token, a heatmap isolating the visual evidence specific to that token from prior-context interference. Applying TAM to a curated set of paintings spanning multiple periods and genres, we analyze grounding patterns across five semantically distinct token categories: common visual objects, style descriptors,

metadata, iconographic tokens, and affective expressions. We find that visual grounding varies substantially with token semantics. We further show that MLLMs attempt to identify artworks and artists, achieving higher accuracy in artist attribution than in title prediction, where hallucinations are more frequent. Finally, we compare TAM with SAM 3 open-vocabulary segmentation. To ensure reproducibility, we release our code, experimental configurations, prompts, and qualitative results on the project page at https://nicolafan.github.io/tamart/.

Keywords: Multimodal Large Language Models · Explainable AI · Art understanding

1 Introduction

Multimodal Large Language Models (MLLMs) [37] have significantly advanced the state of the art in image understanding. These models can describe visual scenes, answer open-ended questions, and perform multimodal reasoning with high degree of fluency. Architectures such as LLaVA [26], Qwen3-VL [5], and InternVL [13] integrate visual encoders with autoregressive language models, allowing them to generate detailed natural-language interpretations of images. Beyond generic visual tasks, these models are increasingly applied and explored in contexts related to cultural heritage, where they demonstrate promising capabilities in artwork captioning, stylistic analysis, and visual question answering.

Artworks pose challenges that differ substantially from photorealistic imagery: understanding a painting requires not only recognition of concrete visual entities, but also interpretation of style, iconographic symbols, and affective atmosphere [11]. A single caption may simultaneously refer to localized objects (e.g., "a woman holding a child"), global stylistic properties (e.g., "loose impressionistic brushwork"), emotional tone (e.g., "a melancholic atmosphere"), and authorship metadata (e.g., "painted by Vincent van Gogh"). Despite the fluency of modern MLLMs, the visual grounding underlying their outputs remains poorly understood—it is often unclear whether a generated claim is anchored in relevant image regions or merely reflects language priors acquired during training. This limitation is particularly critical in cultural heritage applications, where interpretability and trustworthiness are essential.

To address this issue, a large body of work has investigated explainability in deep vision models through methods such as CAM [38], Grad-CAM [33], attention rollout [1], and concept-based attribution techniques such as TCAV [23]. Within the art domain, several studies have applied explainability methods to analyze visual models for tasks such as emotion recognition and multi-task artwork classification [10,31]. However, these approaches typically focus on classification outputs and provide a single global explanation per image. Consequently, they do not capture the token-by-token nature of reasoning in autoregressive multimodal generation.

To overcome this limitation, the Token Activation Map (TAM) approach [24] was introduced to isolate token-specific visual evidence while mitigating interfer-

ence from prior generation context. In this work, we leverage TAM to investigate how MLLMs visually ground different semantic components of artwork descriptions. We apply this framework to captions generated for paintings spanning multiple artistic periods and genres, and analyze the spatial characteristics of activation maps across five semantically distinct categories (Fig. 1). Our goal is to understand whether different aspects of artwork interpretation correspond to distinct patterns of visual grounding, and whether abstract concepts such as style and affect rely more on dispersed visual cues and patterns than concrete depicted entities.

The rest of this paper is structured as follows. Section 2 discusses related work. Section 3 describes the proposed methodology. Section 4 presents experimental results and evaluation. Section 5 concludes the paper and outlines directions for future research.

2 Related Work

2.1 Vision-Language Models for Art Understanding

Interest in applying vision-language models to art understanding predates the MLLM era. Early work on semantic art retrieval [19] and multimodal visual question answering about artworks [20] established the core benchmarks for the field. Contrastive VLMs such as CLIP were subsequently probed on art-domain tasks [31], showing effective stylistic clustering [14] and sensitivity to high-level formal categories such as Wölfflin's principles [21], while also revealing limited perception of period and aesthetic nuance [4].

The emergence of MLLMs has significantly expanded the possibilities for art understanding. Generalist systems such as LLaVA [26], Qwen3-VL [5], and InternVL [13] provide strong baselines, and several works have adapted or evaluated them specifically in the art domain. GalleryGPT [8] fine-tuned LLaVA on formal analysis annotations and exposed systematic failure modes in GPT-4V on canonical paintings. VQArt-Bench [3] constructed an evaluation suite targeting iconographic and symbolic reasoning, revealing significant gaps between proprietary and open-source MLLMs. On the generative side, I Dream My Painting [18] connects MLLMs with diffusion-based inpainting for artwork restoration, NADA [30] harnesses cross-attention maps for annotation-free iconographic detection, and ArtSeek [17] integrates late-interaction retrieval with Qwen2.5-VL for grounded, knowledge-augmented artwork description.

Despite this growing body of work, none of these approaches examines *which visual regions* an MLLM attends to when generating different aspects of an artwork description, which is the central question addressed in this paper.

2.2 Explainability of Vision and Multimodal Models

Interpretability of deep vision models has evolved from gradient-weighted class activation maps [33] and concept-based probing [23] to attention rollout [1] and

relevance-propagation methods for ViTs [12]. For contrastive vision-language models, multi-modal attribution [36] produces joint image-text relevance maps, Visual-TCAV [2] extends integrated-gradient attribution to per-concept localization, and SpLiCE [7] decomposes CLIP embeddings into sparse semantic concepts without task-specific training. Interpreting autoregressive MLLMs is harder, since each output token is generated conditioned on all prior context, making naive activation maps noisy and confounded. The logit-lens family partially addresses this by projecting intermediate activations into the vocabulary space [6], and subsequent probing work finds that visual-to-class projection concentrates in specific mid-to-late LLM layers [28]. TAM [24] resolves the confounding issue more directly: for each generated token it estimates an interference map from earlier context via least-squares subtraction, then applies a rank Gaussian filter to suppress residual noise, yielding a per-token heatmap that localizes the image regions most relevant to each generated word. Evaluated across seven MLLM architectures, TAM establishes state-of-the-art localization performance and provides the methodological basis for the present work.

2.3 Explainability Applied to the Analysis of Artworks

Despite the wide range of existing interpretability methods, their application to the art domain remains sparse. In the context of CNN-based classification, foundational studies using Grad-CAM and its variants [29] directly compared CAM, Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++ on a neural network trained on Christian iconography, showing that the maps highlight iconographically meaningful regions—attributes such as halos, lambs, and arrows—with bounding boxes reaching 55% mean Intersection-over-Union (IoU) against manually annotated iconographic elements. The ArtDL dataset and accompanying ResNet baseline [27] similarly confirmed through CAM visualizations that the model focuses on expected iconographic attributes for individual saint classes. Shifting to transformer architectures, Diem and Mandl [15] contrasted Grad-CAM on a ResNet with attention maps on a ViT for artist-attribution of portrait prints, finding that the ViT attends almost exclusively to the depicted person while the CNN concentrates on local printing-technique textures. Strafforello et al. [34] conducted the first evaluation of generative VLMs on art style, author, and period classification in a zero-shot regime. Most directly related to the present work, Schneider [32] benchmarks seven XAI methods on CLIP in art-historical contexts and finds that the legibility of visual reasoning depends heavily on the conceptual stability and representational availability of the examined categories, while Limpijankit et al. [25] decompose the latent space of VLMs to identify the concepts driving art style prediction, reporting that 73% of extracted concepts were judged coherent by art historians. Both works, however, are anchored in contrastive VLMs and operate at the level of a single classification output rather than token-by-token generation of a full textual description.

Our work takes a complementary step: by applying TAM to a generative MLLM, we obtain a distinct attribution map for *each* word in the painting

description, enabling a systematic comparison of the visual grounding of style-related, content-related, and iconographic tokens, and asking whether an MLLM partitions its visual attention across these semantically distinct facets of art in a way that aligns with art-historical reasoning.

3 Methodology

We describe a pipeline that, given a digitized artwork, produces a set of semantically typed visual activation maps revealing *where* in the image an MLLM grounds different aspects of its generated description—concrete objects, iconographic subjects, stylistic attributes, affective judgments, and metadata. The pipeline proceeds in four stages: we first describe the data, model, and caption generation procedure (Section 3.1); we then summarize the Token Activation Map (TAM) method we adopt for token-level visual grounding (Section 3.2); we classify the generated captions into semantically typed spans (Section 3.3); and finally we aggregate the token-level maps into span-level maps that serve as the basic unit of analysis (Section 3.4).

3.1 Data, Model, and Caption Generation

To study how an MLLM grounds its description of a painting, we assembled a corpus by collecting the 1,000 most-viewed paintings from WikiArt[3] through its public API, downloaded at the platform's maximum resolution. Popularity is a useful proxy here: the most-viewed works skew toward famous paintings, enabling us to probe—through the META category introduced in Section 3.3—whether the model recovers correct metadata for recognizable works, and whether its visual grounding shifts between familiar and unfamiliar ones.

We paired this corpus with Qwen2-VL-2B-Instruct [35] as our MLLM backbone. An MLLM couples a Vision Transformer [16] encoder to an autoregressive LLM through a lightweight projector: the encoder turns the image into patch-level visual tokens, the projector maps them into the LLM's embedding space, and they are concatenated with the tokenized prompt before generation proceeds token by token. What matters for our purposes is that, at every generation step, the hidden state at the last layer aggregates information from the image, the prompt, and all previously generated tokens—the property that the method of Section 3.2 will exploit to recover *where* in the image each generated token is grounded. Qwen2-VL is well-suited to this setting because it preserves fine-grained spatial information more carefully than most MLLMs: it processes images at their native resolution via *Naive Dynamic Resolution* rather than rescaling to a fixed size, and it injects spatial structure directly into attention through 2D Rotary Position Embeddings [22]. In practice, each image is rescaled before encoding to fall within a range of 256 to 1,280 visual tokens, at one token per 28×28 pixel patch, so that the patch grid—to which our activation maps

[3] `https://www.wikiart.org`

will ultimately be aligned—remains faithful to the original image. Its 2B parameter scale also offers a practical balance between representational capacity and computational cost. Furthermore, Qwen2-VL-2B-Instruct is the primary model on which TAM is developed and benchmarked [24], and Li et al. show that the method yields consistent improvements across architecturally distinct MLLMs—including LLaVA-1.5 and InternVL2.5 [24]—with no evidence of qualitatively different grounding behaviour between families; it thus serves as a reasonable proxy for the broader class of decoder-based multimodal models, with extension to further architectures left as future work.

With the corpus and the model in place, generating descriptions is straightforward. For each painting $\mathbf{I}$, we prompt the model with ‘‘`Describe the content and style of this image.`’’ and let it autoregressively produce an answer sequence $t_1^a, \ldots, t_{n_a}^a$ of up to 256 tokens. The prompt is intentionally open-ended: by asking for both *content* and *style*, it invites the model to range over concrete visual objects, stylistic observations, affective impressions, and—for recognized paintings—iconographic references and metadata such as artist or title. As the model generates, we extract a per-token visual activation map for every t_i^a using the procedure described next.

3.2 Token-Level Visual Grounding with TAM

To localize the contribution of each generated token to the image, we adopt the Token Activation Map method of Li et al. [24]. Classical activation-mapping techniques such as CAM [38] and Grad-CAM [33] target classifiers that emit a single prediction, and applying them token-by-token to an autoregressive MLLM yields noisy and unreliable maps. TAM is designed precisely for this setting, and we summarize here only the intuitions needed to follow Section 3.4; we refer the reader to the original paper for the full formulation.

The starting point is a *raw* activation map for each generated token t_i^a, obtained by projecting the visual patch features against the row of the LLM's output classifier corresponding to that token. Intuitively, this measures how strongly each image patch aligns with the concept the token encodes—the same principle as CAM, applied independently at every generation step. The difficulty is that these raw maps are systematically corrupted. Because generation is autoregressive, the hidden state that produces t_i^a has already absorbed information from the image, the prompt, and every earlier generated token, so the raw map for t_i^a inherits *interference* from correlated context tokens that activate similar image regions. If a caption first mentions a *horse* and then names its *rider*, the raw map for “rider” tends to bleed onto the body of the horse, since the two entities are spatially adjacent and the hidden state for “rider” still carries information about the earlier token.

TAM removes this interference by treating each earlier token's raw map as a potential confounder and subtracting an estimated interference map—a relevance-weighted combination of the context maps. The weights are computed from textual similarity between the context tokens and t_i^a; identical tokens are excluded so the model does not suppress genuine self-activation. The optimal

subtraction scale is fit in closed form by least squares. A rank Gaussian filter is then applied to suppress residual salt-and-pepper noise, yielding for each generated token t_i^a a refined visual activation map $\bar{\mathbf{A}}_i^a \in [0,1]^{n_v}$ defined over the n_v patches of the visual encoder's grid. This is the only TAM output we use in what follows: by construction, $\bar{\mathbf{A}}_i^a$ encodes *where* the model grounds token t_i^a in the image, with the contribution of correlated context tokens already removed.

3.3 Semantic Span Classification

An MLLM description of a painting interleaves tokens of very different semantic nature: a single sentence may name a depicted figure, comment on a color palette, ascribe a mood, and recall the artist's name. To disentangle these, we classify contiguous token spans of the generated caption into five categories:

- **CVO** (Concrete Visual Object): a depicted entity or its visible attribute that can be localized in the image, e.g. *"woman"*, *"body of water"*;
- **ICON** (Iconographic Subject): a named mythological, religious, or historical figure depicted in the scene, e.g., *"Dante"*, *"Madonna"*;
- **STYLE** (Painterly Attribute): a reference to brushwork, palette, technique, or period/movement, e.g., *"loose brushwork"*, *"Renaissance"*;
- **AFFECT** (Affective/Interpretive): a mood, emotion, or atmospheric quality, e.g., *"dramatic atmosphere"*, *"serene"*;
- **META** (Metadata): artist name, painting title, date, or provenance, e.g., *"Leonardo da Vinci"*, *"1642"*.

Classification is performed by a separate text-only instruct LLM, Qwen3-4B-Instruct-2507 [5], which we also reuse as an LLM-as-a-judge in Section 4.2. The caption is presented to the classifier as an indexed token list—each token prefixed by its position in the generated sequence—and the model is prompted to return a structured JSON response specifying, for each detected expression, its surface form, the span of token indices $\{j, \ldots, j+l-1\}$ it occupies (where l denotes the expression's length in tokens), and its assigned category. The indexed presentation lets the model report token positions directly, sidestepping the need to align surface forms back to the tokenized sequence post hoc. We cap each expression at ten tokens and discard purely positional fillers (e.g., *"in the foreground"*) to focus on contentful spans. This yields a set of typed spans $\mathcal{S} = \{(e_m, \mathcal{I}_m, y_m)\}_{m=1}^{M}$, where e_m is the expression text, $\mathcal{I}_m \subseteq [1, n_a]$ is the set of token indices, and $y_m \in \{\text{CVO}, \text{ICON}, \text{STYLE}, \text{AFFECT}, \text{META}\}$ is the category. Applied to the full corpus, this procedure produces 12,878 labeled spans, which—together with the per-token activation maps—are precomputed once and reused across all experiments.

3.4 Span-Level Map Aggregation

TAM produces one activation map per token, but many of the semantic concepts identified in Section 3.3 span multiple tokens—*loose brushwork* and *Leonardo da*

Vinci, for instance, are each tokenized into several sub-word units. Li et al. [24] address this for evaluation purposes by selecting, among the tokens of a multi-token word, the one whose map achieves the highest IoU with a ground-truth segmentation mask. In our setting, no such ground truth is available, so an IoU-based selection criterion is inapplicable.

We instead define the span map as the *element-wise average* of the refined per-token maps over the span's constituent tokens:

$$\bar{\mathbf{A}}_m^{\text{span}} = \frac{1}{|\mathcal{I}_m|} \sum_{i \in \mathcal{I}_m} \bar{\mathbf{A}}_i^a. \tag{1}$$

We compared this averaging strategy against two natural alternatives—taking only the first token's map, and taking the per-element maximum across the span—through an informal qualitative inspection on a subset of our captions. Averaging produced the most spatially coherent maps in our setting: first-token selection can under-represent the span's grounding when the semantically distinctive sub-word appears later (a frequent pattern in BPE-style tokenization), while the per-element maximum tends to over-extend activations by inheriting spurious peaks from individual tokens.

The resulting span map $\bar{\mathbf{A}}_m^{\text{span}}$, paired with its category label y_m, is the basic unit of analysis in the experiments of Section 4: it lets us inspect, for instance, whether STYLE spans activate diffusely across the canvas while CVO spans concentrate on localized regions, or whether AFFECT tokens are grounded in specific visual elements rather than responding to the image holistically.

4 Experiments

We evaluated the proposed pipeline by analyzing semantic variation in visual grounding, metadata reliability, and relation to another localization method.

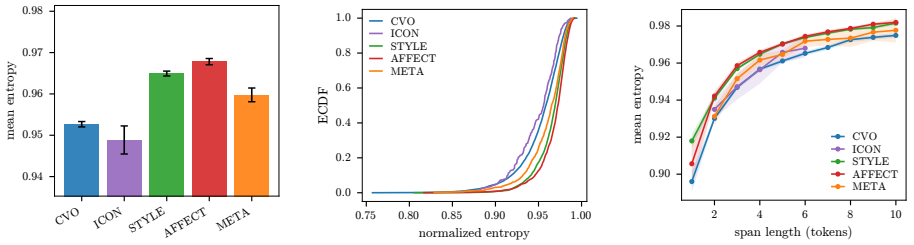
4.1 Spatial Localization of Activation Maps by Content Type

We investigated whether Token Activation Maps localize differently depending on the *type* of content a token describes. For every span, we averaged the TAMs of its tokens and measured the *normalized spatial entropy* of the resulting map—the Shannon entropy of the map, viewed as a distribution over the N vision-grid cells, divided by $\log N$—which lies in $[0, 1]$, with 1 fully diffuse and 0 fully localized. We complemented entropy with two concentration measures: the Gini coefficient of the map values and the fraction of activation in the hottest 10% of cells.

Concrete content is significantly more localized than abstract content (Table 1). Mean entropy is lowest for CVO (0.953) and ICON (0.949) and highest for STYLE (0.965) and AFFECT (0.968), with META (0.960) in between; the Gini coefficient and top-10% mass ranked the categories in exactly the opposite order, confirming the metric. Figure 2a reports the per-type means with 95% confidence intervals and Figure 2b the cumulative distributions, which are cleanly

Table 1: Localization of token activation maps by content type (N = 12,878 spans). Normalized spatial entropy, the Gini coefficient, and top-10% mass concentration. Arrows indicate the direction of stronger localization; cells are shaded so that darker indicates greater localization. Concrete content (CVO, ICON) is the most localized, abstract content (STYLE, AFFECT) is the most diffuse.

Span type	n	Entropy ↓	Gini ↑	Top-10% ↑
CVO	6241	0.953 ± 0.026	0.421	0.303
ICON	197	0.949 ± 0.024	0.443	0.312
STYLE	3542	0.965 ± 0.018	0.368	0.264
AFFECT	2172	0.968 ± 0.018	0.350	0.259
META	726	0.960 ± 0.023	0.392	0.282



(a) (b) (c)

Fig. 2: Activation maps localize by content type. (a) Mean normalized spatial entropy per span type (95% CI); lower is more localized. (b) Cumulative distributions of entropy. (c) Mean entropy vs. exact span length: the ordering holds at every length, so it is not an artifact of token-averaging.

stochastically ordered. A Kruskal–Wallis test rejected equality of the distributions ($p < 0.001$), and the effect survived controlling for within-painting dependence (paired Friedman test over the paintings containing every category, $p < 0.001$).

Because a span map is the mean of its token maps, longer spans average more maps and are mechanically biased toward higher entropy, and span length correlates with both category and entropy (Spearman $\rho \approx 0.72$). We therefore confirmed that the effect is not a length artifact. Figure 2c plots mean entropy against the *exact* span length: the per-type curves stayed separated and ordered at every length, including single-token spans, for which no averaging occurs. With length held fixed, the categories still differed ($p < 0.001$ within each length), and regressing out the length trend left the ordering intact (Kruskal–Wallis on residuals, $p < 0.001$). Figure 3 shows one representative map per type: CVO and ICON spans concentrate on a single object, whereas STYLE and AFFECT spans draw on the whole canvas.

Fig. 3: Representative token activation maps, one per content type. Concrete objects (CVO) and named icons (ICON) localize tightly, while style (STYLE) and affect (AFFECT) spans are diffuse; META lies in between. Notice how the META span *"The Kiss"* corresponds to a model hallucination.

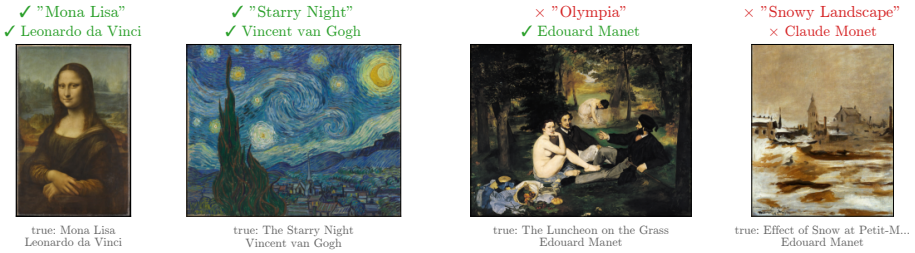
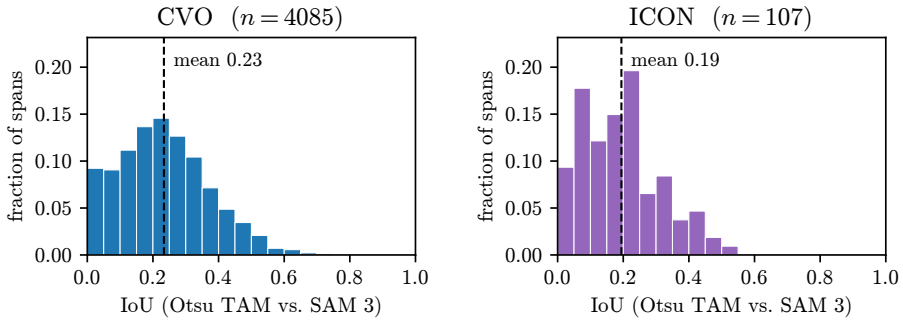


Fig. 4: Title and artist predictions extracted from META spans and graded by an LLM judge (green: correct, red: wrong), with the ground-truth metadata below each painting. Left to right: both correct (*Mona Lisa*, *The Starry Night*); artist correct but title wrong (Manet's *Luncheon on the Grass*, guessed as *Olympia*, another Manet); and both wrong (a Manet snow scene attributed to Monet).

4.2 Title and Artist Prediction in META Spans

A recurring behavior was that some META spans went beyond providing general background about the artwork and explicitly stated the *title* or the *artist*, which, although confidently phrased, often appeared to be guessed from contextual cues. To assess how reliable these descriptions were, we used an LLM-as-a-judge: for every artwork, the judge read its META spans, extracted the predicted title and artist when present, and compared each against the ground-truth metadata, yielding a per-painting correctness verdict for the title and for the artist.

Our results indicated that the model named the *artist* correctly far more often than the *title* (roughly 82% vs. 28%): it more often identified the artist, frequently on the basis of distinctive stylistic cues, while producing a plausible but incorrect title. Figure 4 shows the typical outcomes: two canonical works for which both the title and artist names are correct; one where the artist name is correct but the title corresponds to a different work by the same painter, and one where both title and artist name are incorrect.



(a) (b)

Fig. 5: IoU between the Otsu-thresholded TAM map and the SAM 3 concept mask, per category, over spans SAM 3 localized (relative frequency; dashed line = mean). Agreement is modest and higher for (a) concrete objects than for (b) named iconographic subjects.

We additionally tested whether any TAM statistic of the predicting span anticipated its correctness: where the map concentrated (entropy), how strong it was (mass), and how much the prediction leaned on the image rather than on the already-generated text (its *visual reliance*, the share of evidence drawn from the image vs. the co-text). Map shape carried no signal—neither entropy nor mass separated correct from wrong predictions within either type. The only statistic with any association was visual reliance, and only for the artist: correct artist predictions drew on average 0.93 of their evidence from the image against 0.85 for wrong ones (Mann–Whitney $p \approx 0.002$, Cliff's $\delta \approx 0.30$), i.e., wrong predictions were markedly more text-driven, whereas for titles there was no such effect ($p \approx 0.12$). The association was genuine but weak, and largely a painting-level property: it washed out once visual reliance was measured relative to each artwork's own baseline, indicating it reflected whether the description as a whole stayed visually grounded rather than anything specific to the predicting token. As a correctness classifier it is therefore only a triage signal (ROC-AUC ≈ 0.65), not a reliable predictor. We read it as a hint about *why* the model is sometimes right—it was looking at the canvas rather than confabulating from a language prior—and as evidence that titles tend to be invented from priors while artists can be genuinely recognized from the image.

4.3 TAM as an Open-Vocabulary Detector

The previous experiments treated the activation maps as post hoc explanations. Here we ask a more practical question: can a token activation map be repurposed as an *open-vocabulary detector* for the visual arts? If a single forward pass of the captioning model already yields, for any noun phrase it produces, a spatial map

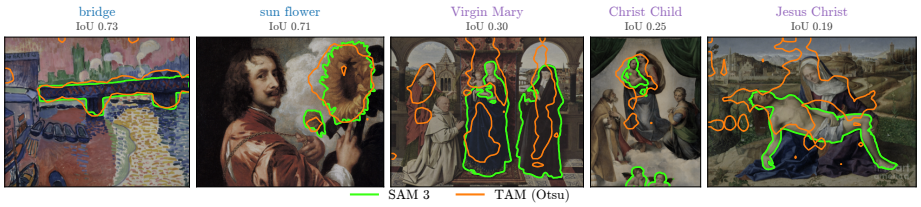


Fig. 6: What each model localizes for a span (caption text above each panel, blue = CVO, purple = ICON; IoU below). The painting is dimmed; the green outline is the SAM 3 concept mask, and the orange outline is the Otsu-thresholded TAM prediction, which is coarser for both categories. The ICON panels show characteristic errors: over-detection of *Virgin Mary*; TAM correctly (although coarsely) focuses on the *Christ Child* while SAM 3 also segments the cherubs; and SAM 3 cleanly segments *Jesus Christ* while TAM only points at the scene center.

that lands on the right region, then localization comes "for free" alongside the description, without a dedicated segmentation model.

We probed this by comparing it with SAM 3 [9], a state-of-the-art promptable concept segmenter used as an open-vocabulary reference. For every CVO (concrete visual object) and ICON (named iconographic subject) span, we prompted SAM 3 with the span's surface text and took the union of its instance masks as the reference region; we binarized the corresponding TAM map with Otsu's threshold and reported the IoU between the two. SAM 3 returned no detection for 35% of CVO and 46% of ICON spans; lacking any reference, these were excluded from the IoU statistics.

Figure 5 reports the IoU distributions. For both categories, the agreement is modest, and it is higher for concrete objects than for named iconographic subjects—CVO mean 0.233 (median 0.224) versus ICON mean 0.194 (median 0.184). What is clear from inspecting the maps is that the Otsu-thresholded TAM predictions are consistently *coarser and less precise* than the SAM 3 masks, for CVO as well as ICON spans: where the two overlap, it is typically a loose TAM blob loosely enclosing a much tighter SAM region. We do not find that either method is systematically better at localizing.

The qualitative examples in Figure 6 make the iconographic error modes concrete. For *Virgin Mary* (third panel), both models over-detect, marking several of the depicted figures as the Virgin rather than one. For *Christ Child* (fourth panel, Raphael's *Sistine Madonna*) TAM—despite its coarse mask—correctly concentrates on the infant the model was attending to, whereas SAM 3, lacking the semantic context, additionally segments the two cherubs at the foot of the composition. For *Jesus Christ* (fifth panel, a *Pietà*), the situation reverses: SAM 3 cleanly segments the figure of Christ, while TAM only indicates that the model is attending to the center of the scene. Since neither approach is uniformly preferable, a promising direction for future work is to combine the se-

mantic grounding of TAM with the precise boundaries of a dedicated segmenter to obtain better localizations than either alone.

5 Conclusion

In this work, we studied how MLLMs visually ground different semantic components of artwork descriptions using TAM as our interpretability lens.

Several directions remain open. Extending the analysis to larger and architecturally diverse MLLMs would clarify how robust the grounding patterns are across the design space. The hallucination behavior observed in the title prediction calls for targeted mitigation strategies, such as retrieval-augmented generation or uncertainty-aware decoding, whose effectiveness could be directly measured within the TAM framework. Fusing TAM with precise segmentation models offers a promising route toward fine-grained, annotation-free iconographic localization—a practically valuable capability for digital humanities and cultural heritage applications. Finally, scaling the dataset to broader artistic traditions and underrepresented periods would test whether the grounding patterns identified here generalize beyond the Western canon.

Acknowledgements. Nicola Fanelli's research is funded by a Ph.D. fellowship under the Italian "D.M. n. 118/23" (NRRP, Mission 4, Investment 4.1, CUP H91I23000690007).

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 4190–4197 (2020)
2. Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From attribution maps to human-understandable explanations through concept relevance propagation. Nature machine intelligence **5**(9), 1006–1019 (2023)
3. Alfarano, A., Venturoli, L., Del Castillo, D.N.: VQArt-Bench: A semantically rich VQA Benchmark for Art and Cultural Heritage. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 396–406 (2025)
4. Asperti, A., Dessi, L., Tonetti, M.C., Wu, N.: Does CLIP perceive art the same way we do? In: 2025 International Conference on Content-Based Multimedia Indexing (CBMI). pp. 1–8. IEEE (2025)
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
6. Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., Biderman, S., Steinhardt, J.: Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112 (2023)
7. Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F.P., Lakkaraju, H.: Interpreting clip with sparse linear concept embeddings (splice). Advances in Neural Information Processing Systems **37**, 84298–84328 (2024)

8. Bin, Y., Shi, W., Ding, Y., Hu, Z., Wang, Z., Yang, Y., Ng, S.K., Shen, H.T.: Gallerygpt: Analyzing paintings with large multimodal models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7734–7743 (2024)
9. Carion, N., et al.: SAM 3: Segment anything with concepts. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=r35clVtGzw
10. Castellano, G., Miccoli, M.G., Scaringi, R., Vessio, G., Zaza, G., et al.: Using LLMs to explain AI-generated art classification via Grad-CAM heatmaps. In: XAI. it@ AI* IA. pp. 65–74 (2024)
11. Cetinic, E., Lipic, T., Grgic, S.: A deep learning perspective on beauty, sentiment, and remembrance of art. IEEE access **7**, 73694–73710 (2019)
12. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 782–791 (2021)
13. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
14. Conde, M.V., Turgutlu, K.: Clip-art: Contrastive pre-training for fine-grained art classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3956–3960 (2021)
15. Diem, S., Mandl, T.: Automatic Classification of Portraits: Application of Transformer and CNN Based Models for an Art Historic Dataset. In: LWDA. pp. 192–206 (2023)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (2021)
17. Fanelli, N., Vessio, G., Castellano, G.: ArtSeek: Deep artwork understanding via multimodal in-context reasoning and late interaction retrieval. arXiv preprint arXiv:2507.21917 (2025)
18. Fanelli, N., Vessio, G., Castellano, G.: I dream my painting: Connecting MLLMS and diffusion models via prompt generation for text-guided multi-mask inpainting. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6073–6082. IEEE (2025)
19. Garcia, N., Vogiatzis, G.: How to read paintings: semantic art understanding with multi-modal retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
20. Garcia, N., Ye, C., Liu, Z., Hu, Q., Otani, M., Chu, C., Nakashima, Y., Mitamura, T.: A dataset and baselines for visual question answering on art. In: European conference on computer vision. pp. 92–108. Springer (2020)
21. Ghildyal, A., Wang, L.Y., Liu, F.: WP-CLIP: Leveraging CLIP to Predict Wolfflin's Principles in Visual Art. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 396–405 (2025)
22. Heo, B., Park, S., Han, D., Yun, S.: Rotary position embedding for vision transformer. In: European Conference on Computer Vision. pp. 289–305. Springer (2024)
23. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)

24. Li, Y., Wang, H., Ding, X., Wang, H., Li, X.: Token activation map to visually explain multimodal llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 48–58 (2025)
25. Limpijankit, M., Alshomary, M., Daoud, Y.O., Ananthram, A., Trombley, T., Spratt, E.L., Filonenko, A., Pivo, H., Stengel-Eskin, E., Bansal, M., et al.: Does AI See like Art Historians? Interpreting How Vision Language Models Recognize Artistic Style. arXiv preprint arXiv:2603.11024 (2026)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
27. Milani, F., Fraternali, P.: A dataset and a convolutional model for iconography classification in paintings. Journal on Computing and Cultural Heritage (JOCCH) **14**(4), 1–18 (2021)
28. Neo, C., Ong, L., Torr, P., Geva, M., Krueger, D., Barez, F.: Towards interpreting visual information processing in vision-language models. In: International Conference on Learning Representations. vol. 2025, pp. 57172–57189 (2025)
29. Pinciroli Vago, N.O., Milani, F., Fraternali, P., da Silva Torres, R.: Comparing cam algorithms for the identification of salient image features in iconography artwork analysis. Journal of Imaging **7**(7), 106 (2021)
30. Ramos, P., Gonthier, N., Khan, S., Nakashima, Y., Garcia, N.: No annotations for object detection in art through stable diffusion. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6228–6237. IEEE (2025)
31. Scaringi, R., Fiameni, G., Vessio, G., Castellano, G.: GraphCLIP: Image-graph contrastive learning for multimodal artwork classification. Knowledge-Based Systems **310**, 112857 (2025)
32. Schneider, S.: On the Explainability of Vision-Language Models in Art History. arXiv preprint arXiv:2602.20853 (2026)
33. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
34. Strafforello, O., Soydaner, D., Willems, M., Maerten, A.S., De Winter, S.: Have large vision-language models mastered art history? In: International Conference on Image Analysis and Processing. pp. 524–544. Springer (2025)
35. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
36. Wang, Y., Rudner, T.G., Wilson, A.G.: Visual explanations of image-text representations via multi-modal information bottleneck attribution. Advances in Neural Information Processing Systems **36**, 16009–16027 (2023)
37. Wu, J., Gan, W., Chen, Z., Wan, S., Yu, P.S.: Multimodal large language models: A survey. In: 2023 IEEE International Conference on Big Data (BigData). pp. 2247–2256. IEEE (2023)
38. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)