

# Art through CLIP’s Eyes

ANDREA ASPERTI, LEONARDO DESSI, and NICO WU, Department of Informatics: Science and Engineering (DISI), University of Bologna, Bologna, Italy

Despite significant advances in text-based reasoning, the ability of multimodal models to perceive and interpret visual art remains poorly understood. This study explores how CLIP “sees” and understands artworks—focusing on European paintings from the Renaissance onward—by examining its capacity to recognize style, historical period, and cultural context. We evaluate both the strengths and limitations of CLIP’s visual perception, particularly its alignment with human judgments. Our analysis spans paintings created by both humans and AI, probing CLIP’s sensitivity to subtle yet meaningful differences between them, as well as its ability to detect characteristic artifacts and distortions introduced by generative models. The investigation draws on several recent datasets of AI-generated art, offering a critical assessment of their data acquisition methodologies, structure, and benchmarking potential. This work contributes to bridging computational perception and human artistic interpretation, offering new insights into the evolving role of AI in art analysis.

CCS Concepts: • **Applied computing** → **Arts and humanities**; • **Computing methodologies** → **Computer vision**;

Additional Key Words and Phrases: CLIP, multimodal models, style analysis, vision-language alignment, computational art, visual perception

## ACM Reference format:

Andrea Asperti, Leonardo Dessi, and Nico Wu. 2026. Art through CLIP’s Eyes. *ACM J. Comput. Cult. Herit.* 19, 2, Article 32 (June 2026), 33 pages.  
<https://doi.org/10.1145/3812546>

## 1 Introduction

Multimodal models are reshaping how machines perceive and interpret the world by integrating visual and textual information. A prominent example is **Contrastive Language–Image Pretraining (CLIP)** [37], which has become a cornerstone of modern vision–language research.

Trained on hundreds of millions of image–text pairs, CLIP exhibits strong performance across a wide range of tasks, including zero-shot classification [24, 46, 49], image retrieval [8, 29, 48], re-identification [5, 25], prompt-based generation [20, 43], and semantic search [30, 36, 55]. Its ability to align visual and linguistic representations has made it integral to state-of-the-art generative models and vision–language pipelines. In particular, CLIP serves as a guiding mechanism [28, 44] in text-to-image generation systems such as GLIDE [33], DALL-E [38], and Stable Diffusion [15, 39], as well as in creative applications such as style transfer and visual storytelling.

Authors’ Contact Information: Andrea Asperti (corresponding author), Department of Informatics: Science and Engineering (DISI), University of Bologna, Bologna, Italy; e-mail: andrea.asperti@unibo.it; Leonardo Dessi, Department of Informatics: Science and Engineering (DISI), University of Bologna, Bologna, Italy; e-mail: leonardo.dessi@studio.unibo.it; Nico Wu, Department of Informatics: Science and Engineering (DISI), University of Bologna, Bologna, Italy; e-mail: nico.wu@studio.unibo.it.



This work is licensed under Creative Commons Attribution International 4.0.

© 2026 Copyright held by the owner/author(s).

ACM 1556-4711/2026/6-ART32

<https://doi.org/10.1145/3812546>

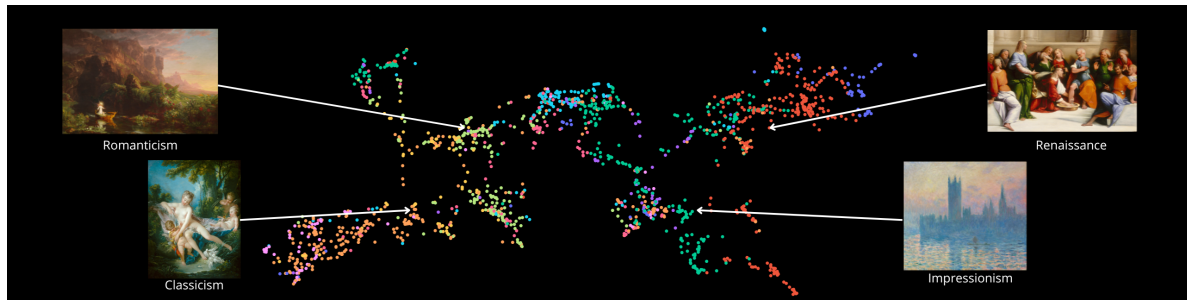


Fig. 1. 3D UMAP projection of image embeddings of National Gallery of Art Dataset extracted from the CLIP ViT-L/14 model. Each point is a different painting, colored by artistic style. UMAP, Uniform Manifold Approximation and Projection.

Despite its widespread adoption, the nature and limitations of CLIP’s perceptual alignment remain insufficiently understood. Although CLIP is trained to match images with descriptive captions, it is unclear to what extent this alignment reflects human perception, especially in complex, subjective, or culturally embedded domains such as visual art. Its ability to identify what is depicted is well documented, but its understanding of how content is rendered through style, composition, or historical context remains largely unexplored. Moreover, existing evaluations tend to focus on natural images and benchmark datasets, leaving open the question of how CLIP performs on ambiguous, interpretive, or aesthetic content such as paintings, particularly those created by AI.

This article investigates CLIP’s perceptual capabilities in the context of visual art, examining both human-made and AI-generated paintings.

We treat CLIP’s vision encoder as a fixed perceptual system—without modifications or fine-tuning—analogously to the human sensory apparatus. This choice is intentional: our objective is to determine whether stylistic, temporal, and qualitative cues relevant to artistic interpretation are already present in CLIP’s latent visual representations, or whether they only emerge through task-specific supervision. Rather than aiming to optimize performance or establish a new benchmark, our contribution is therefore diagnostic and interpretive: we seek to characterize how CLIP organizes and prioritizes visual information when confronted with artworks. A visualization of data in CLIP’s latent space using the **Uniform Manifold Approximation and Projection (UMAP)** algorithm [32] can be seen in Figure 1.

Our goal is to assess CLIP’s ability to capture not only semantic content but also stylistic characteristics, temporal cues, and visual distortions. Through a series of probing tasks and targeted analyses, we evaluate CLIP’s internal representations along multiple interpretive dimensions, including scene type, artistic style, historical period, and the presence of visual artifacts. From this perspective, we argue that progress in art-related vision–language tasks may benefit less from repeatedly fine-tuning the visual encoder, and more from modular and configurable mechanisms—such as guided projections or subspace selection—that selectively exploit features already encoded in the representation space.

To support our probing investigation, we rely on several recent, richly annotated datasets: the National Gallery of Art in Washington [34], the AI-ArtBench dataset [26], and AI-WikiArt [16], and [4] collections.

The inclusion of both human-made and AI-generated artworks is a distinctive aspect of our study. This dual perspective allows us to assess the extent to which models like CLIP can recognize the subtle—but for human observers often evident—stylistic artifacts, inconsistencies, and biases introduced by generative processes. By comparing how CLIP processes real and synthetic paintings, we probe its sensitivity to the visual imperfections and aesthetic divergences that distinguish artificially produced images from their historical counterparts.

This study is part of a broader and ambitious project that seeks to determine whether large vision–language models can develop an aesthetic sense. Do they internalize abstract concepts such as style, harmony, or beauty—and if so, are these grounded in visual abstraction, learned statistical regularities, or biases within their training data?



Since the understanding of aesthetics is fundamentally rooted in direct sensory experience, it is natural to ask how well modern vision systems can support a model's capacity for aesthetic perception.

We approach these questions through both conceptual reflection and empirical analysis. Our results point to a persistent gap between CLIP's textual associations and its sensitivity to visual nuance. Although the model excels at high-level semantic alignment, it consistently struggles with stylistic detail, period attribution, and the identification of visual artifacts in AI-generated imagery. These shortcomings suggest a more fundamental limitation: despite its impressive multimodal capabilities, CLIP lacks a coherent internal model of aesthetic form.

Through a combination of metadata-based evaluation and perceptual baselines, we offer a structured critique of CLIP's performance in aesthetic contexts. In doing so, we emphasize the need for more interpretable and perceptually aligned multimodal systems, especially as models like CLIP increasingly play a role in guiding, evaluating, and curating creative content. As such systems become embedded in artistic workflows and cultural infrastructures, the critical question becomes not only what they can recognize or produce, but how they see, and whether that mode of seeing aligns with human perception.

## 2 Related Works

We divide the discussion of related works in three subsections: CLIP's limitations and robustness (Section 2.1), CLIP's improvements (Section 2.2), and art-related works (Section 2.3).

### 2.1 CLIP's Limitations and Robustness

Several recent studies have investigated the capabilities and limitations of CLIP's visual processing, particularly focusing on how its image embeddings can be accessed and meaningfully exploited through textual prompts. A recurring concern is whether CLIP truly grasps the semantic depth of its prompts. For instance, Pan et al. [35] question the extent to which CLIP understands textual meaning beyond surface-level associations. Recent quantitative benchmarks support this, showing that CLIP demonstrates a strong concept association bias, where it prioritizes frequent co-occurrences over actual visual evidence [47], and struggles significantly with logical operators such as negation [2].

A detailed evaluation of CLIP's behavior in multi-object scenes is presented in [1], where controlled experiments reveal notable biases: the image encoder tends to favor larger objects, while the text encoder gives priority to items mentioned earlier in the prompt. These findings highlight structural asymmetries in the alignment process between modalities. These limitations are further quantified by a documented quantity bias, where the model's failure rates in object counting tasks increase significantly as scene complexity grows [52].

CLIP's robustness has also been a topic of investigation. In [40], the authors assess its vulnerability to adversarial attacks in the context of AI-generated image detection, showing that CLIP-based methods are susceptible to white-box manipulations. A broader evaluation is provided in [45], where CLIP models, despite certain weaknesses, are shown to outperform traditional ImageNet-based vision models in terms of robustness to variations in visual factors.

Another line of inquiry focuses on understanding CLIP's internal representations through inversion techniques. In [22], traditional gradient ascent methods [14, 31, 50] are employed to reconstruct images from CLIP embeddings, offering insight into what the model "sees."

### 2.2 CLIP's Improvements

In light of CLIP's known limitations, a growing body of research has explored methods to improve its performance, enhance downstream generalization, and reduce the modality gap between visual and textual representations. One widely adopted strategy involves the use of adapters—lightweight modules inserted into the model architecture that enable efficient fine-tuning without modifying the core weights. Representative examples include CLIP-Adapter

[17], TIP-Adapter [51], and LlxP [41], which have shown strong performance improvements on vision–language tasks with minimal training overhead.

However, our objective is to probe CLIP’s native perceptual capabilities, especially in the domain of visual art, where interpretability and model transparency are critical. For this reason, we consider adapter-based methods misaligned with the spirit of our investigation: rather than modifying the model for improved performance, we aim to analyze the inductive biases and visual priors encoded through CLIP’s original contrastive pretraining.

Another line of research focuses on learnable prompt embeddings, which are optimized during training to better align textual and visual features. In the case of zero-shot classification, CoOp [54] and CoCoOp [53] extend CLIP’s prompts with trainable vectors, learned via gradient descent. While effective for some tasks, we found this strategy less suitable for the aesthetic domain: the semantic gap between CLIP’s image embeddings and abstract stylistic descriptors in art is often too large to be bridged by prompt tuning alone.

A more promising approach to addressing perceptual misalignments involves projecting CLIP’s latent features into lower-dimensional or task-specific subspaces, as proposed in works like APE and CLIP-Subspace [56, 57]. Since cosine similarity is not invariant under linear transformation, applying projection layers can shift emphasis toward features more relevant to a given interpretive task, such as style or period classification. This technique retains the pretrained encoder while enabling models to focus on interpretively meaningful components of the representation space.

This idea also informs the architecture of recent vision–language models, such as LLaVA [27], Qwen-VL [6], and InstructBLIP [13], which employ learnable projections to map visual features into the embedding space of a **Large Language Model (LLM)**. These projections serve to reduce the modality gap and facilitate instruction-following, without requiring full fine-tuning of the image encoder.

A closely related technique is the use of query embeddings, found in models like BLIP-2 [23], InstructBLIP [13], and ShareGPT4V [11]. These are learnable tokens that interact with the image encoder’s output through cross-attention layers within a separate transformer module, commonly called the Q-Former. This module acts as a visual adapter, extracting task-relevant visual representations that can then be passed into the LLM. Although these models often keep the vision encoder frozen, which partially aligns with our methodology, their focus remains on downstream performance in instruction-following rather than interpretive transparency.

Together, these approaches reflect a growing effort to either enhance CLIP’s downstream utility or to better access and interpret the structure of its embedding space. While we avoid direct fine-tuning or architectural modification, we view techniques such as subspace projections as compatible with our diagnostic goals, insofar as they reveal how stylistic or historical information may be implicitly represented. In what follows, we focus on evaluating CLIP’s native visual representations in the context of art, aiming to characterize their structure, limits, and alignment with human perception.

### 2.3 Art-Related Works

An early attempt to adapt CLIP to the artistic domain was made in CLIPArt [12], where the model was fine-tuned using image–text pairs from the SemArt dataset [18]. Additional datasets relevant to art-focused vision–language tasks include AQUA [19] and ArtQuest [10], both designed for visual question answering. The ArtBench dataset [26] offers a curated collection of art reproductions from museums and archives, specifically created to benchmark style classification performance.

Our approach is related to the investigation presented in [3], which, however, as noted by the authors, reported only preliminary findings. We build on this line of inquiry by extending the analysis to larger and more diverse datasets of artworks, and by conducting a more comprehensive set of experiments within a more systematic evaluation framework.

Another study that closely aligns with the goals of our investigation is GalleryGPT [9]. This work makes two notable contributions. First, it introduces a large dataset of well-known paintings accompanied by detailed

formal analyses, automatically generated using GPT-4 and Gemini. Second, it presents a fine-tuned version of ShareGPT4V, referred to as GalleryGPT, tailored for generating artistic analyses, and demonstrates improved results across several tasks. Notably, the visual encoder in GalleryGPT is kept frozen, which aligns with our methodological stance. However, the increasing availability of expert-authored descriptions from institutions such as the National Gallery of Art diminishes the practical need for LLM-generated commentary. Moreover, despite progress, the overall performance of current systems on complex art analysis remains modest, underscoring the need to critically examine the underlying visual mechanisms, that is the central focus of our work.

A further methodological distinction from GalleryGPT lies in our inclusion of both human-made and AI-generated artworks; this enables a more nuanced evaluation of CLIP's ability to identify subtle stylistic artifacts and distortions introduced by generative processes.

### 3 Datasets

Our investigation is primarily focused on European paintings from the Renaissance onward. The study was conducted on several recent datasets that we shall introduce and discuss in this section: a subset of the freely available National Gallery of Art collection in Washington, the AI-ArtBench dataset [42], AI-WikiArt [16], and AI-Pastiche [4].

#### 3.1 National Gallery of Art Dataset (NGAD)

The NGAD was created by selecting 2,693 artworks from the National Gallery's publicly available collection [34], focusing on European paintings.

Each record in the dataset comprises 11 key attributes, including a unique identifier (`objectId`), title, creation period (`period`), artist or attribution, and a high-resolution image link (`link`) provided via the IIIF protocol. A key feature of the dataset is the description attribute, which provides a detailed textual representation of each artwork. These descriptions are structured to convey visual elements, artistic techniques, and compositional aspects.

Since CLIP processes textual inputs with a limit of 77 tokens, a summarized version of each description was generated to ensure compatibility with the model. These summaries retain essential information regarding the subject, composition, and artistic style while conforming to the token constraint. An example of a full description and its corresponding summary is given in Appendix A. In the same appendix, we also discuss different summarization approaches and their impact on similarity scores.

The dataset covers a broad range of historical periods and artistic styles. The most represented styles include Renaissance (407 artworks), Baroque (395), Realist (353), Naive (288), Impressionism (284), and Post-Impressionist (284).

#### 3.2 AI-ArtBench

The AI-ArtBench dataset is a large-scale collection developed within the ArtBrain project [42] by researchers at Carnegie Mellon University and UC Berkeley, aimed at benchmarking AI-generated artwork and evaluating models for style classification and source attribution (human vs. AI).

It contains 185,015 images annotated across 10 artistic styles, combining:

- 60,000 human-created artworks (from ArtBench-10, with 6,000 images per style), and
- 125,015 AI-generated images produced using Latent Diffusion and Stable Diffusion, conditioned on textual prompts.

The dataset spans styles such as Baroque, Impressionism, Surrealism, Ukiyo-e, and others. Human images are provided at  $256 \times 256$  resolution, while synthetic ones include both  $256 \times 256$  and  $768 \times 768$  versions. ArtBench-10



Fig. 2. Examples of AI-ArtBench-generated images and their nearest neighbor, demonstrating the visual similarity that leads to inflated performance metrics.

is notable for its clean annotations, class balance, and high quality, avoiding the long-tail issues common in earlier art datasets.

While AI-ArtBench’s balanced structure is a notable strength, certain aspects of the dataset raise concerns regarding its suitability for nuanced evaluation. First, the quality of the image generators used—Latent Diffusion and early versions of Stable Diffusion—is relatively modest by current standards. These models may fail to capture the finer stylistic details necessary for credible artistic imitation, especially in more complex or subtle styles.

More critically, the prompting strategy used during generation is extremely minimal: each image is generated using a simple sentence of the form [“A painting in <art style> art style”]. This generic phrasing leaves considerable freedom to the generator, which may respond by producing superficial or stereotypical representations of the target style. As a result, outputs are likely to be very similar to each other, emphasizing easily recognizable, low-complexity visual cues, reinforcing the model’s existing biases or stylistic shortcuts, rather than challenging its representational capacity. This is quite apparent in Figure 2, where we visualize a few generated images along with their five nearest neighbors in the dataset.

This design choice risks skewing the dataset toward paradigmatic and often trivial examples of each style, resulting in images that are not only of low aesthetic quality but also highly stereotypical and repetitive. In some cases, the prompt itself (see for instance the phrase “a painting in Renaissance style” in Figure 2) is visibly embedded within the generated image, undermining its stylistic credibility. The high degree of visual similarity among samples further exacerbates the issue: in a random train/test split, there is a substantial likelihood of nearly identical images appearing across subsets, with the risk increasing in proportion to the dataset size. Such artifacts may inflate model performance and obscure meaningful evaluation of a system’s ability to generalize across stylistic nuance. Using richer, semantically informed prompts—closer to curatorial descriptions—could have introduced greater visual diversity and interpretive depth.



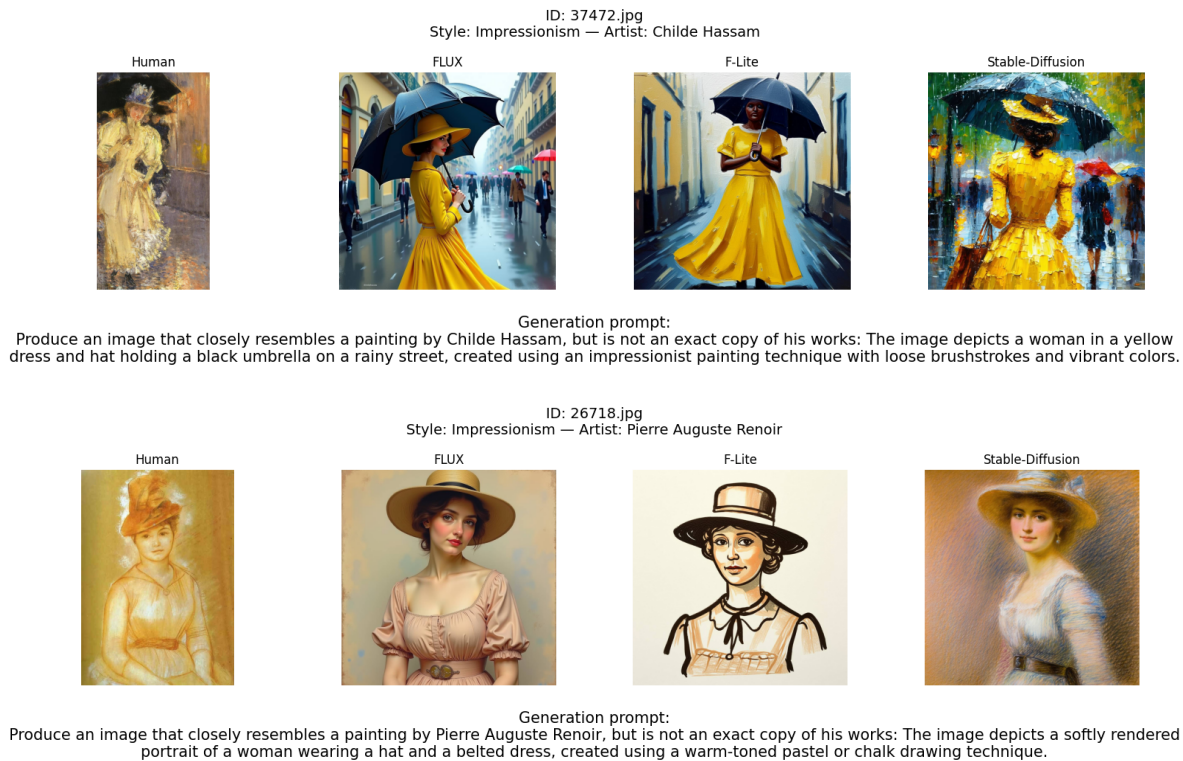


Fig. 3. Examples from the AI-WikiArt dataset. From left to right: the human-generated original image, followed by versions generated by Flux, F-Lite, and Stable Diffusion.

### 3.3 AI-WikiArt

The AI-WikiArt dataset is a large-scale collection designed to support the systematic evaluation of vision-language models in tasks such as authorship attribution and the detection of synthetically generated art [16]. It is distinguished by its dual structure, comprising both authentic and AI-generated artworks.

The real paintings subset includes 39,530 images curated from WikiArt. Only paintings with identifiable artists were retained, excluding works labeled as “unknown.” The resulting dataset spans 128 artists, 10 genres, and 27 artistic styles, with metadata detailing the artist, genre, and style for each work.

To complement this, the dataset includes a synthetic subset composed of AI-generated counterparts designed to imitate the real paintings. These were produced through a two-step automated pipeline: first, descriptive captions for the real images were generated using GPT4.1-mini; second, these captions were used as prompts to synthesize new images using three text-to-image models—Stable Diffusion, Flux, and F-Lite.

The prompts were carefully crafted to instruct the models to “produce an image that closely resembles a painting by [correct painter], but is not an exact copy of his works.” This approach ensured that the synthetic images were aligned with the real dataset in content and style, allowing for fair comparisons in model evaluation.

Examples from the AI-WikiArt dataset are presented in Figure 3. As in the case of AI-ArtBench, the image generators used—Stable Diffusion, Flux, and F-Lite—represent relatively modest models by today’s standards. The resulting images often lack the stylistic richness and formal complexity characteristic of the original paintings, particularly in terms of color, scene composition, and historical nuance. This limitation casts doubt on the feasibility of using such images for tasks like authorship attribution, where subtle stylistic cues are essential.



Fig. 4. Examples of images from the AI-Pastiche dataset. All images have been generated from the following prompt: “Generate a detailed winter landscape painting in the Flemish renaissance style of the second half of the XVI century. Depict a snow-covered village with small, rustic houses nestled into a hilly landscape. Include bare, slender trees in the foreground with hunters walking through the snow, accompanied by dogs. The scene should feature frozen lakes or ponds in the background, where villagers are skating and engaging in winter activities. The sky is a muted, wintry blue-gray, and the overall tone of the painting should evoke a peaceful, yet somewhat melancholic atmosphere, with intricate details showing rural life during winter.”

Compared to AI-ArtBench, the prompt design in AI-WikiArt is sensibly more sophisticated, as it includes artist-specific references and encourages stylistic imitation. However, even these prompts remain relatively constrained. With the capabilities of modern generative models, far more detailed and semantically rich descriptions could be leveraged to produce outputs that better approximate the visual and stylistic qualities of real artworks. Incorporating such detail would not only improve the quality of the synthetic data but also enable more meaningful evaluation of visual and stylistic perception in downstream tasks.

### 3.4 AI-Pastiche

The AI-Pastiche dataset [4] is a curated collection of 953 AI-generated paintings designed to mimic historical artistic styles. Created using 73 meticulously crafted prompts, the dataset spans a wide range of art periods. The images were generated using multiple state-of-the-art generative models and are intended to support the evaluation of how convincingly AI can reproduce the visual characteristics of human art. Examples of images and the corresponding prompts are given in Figure 4.

Each image is accompanied by rich metadata, including the generative model used, the prompt text, intended style and period, as well as a list of subject descriptors like “crowd,” “landscape,” or “soft tones.”

A distinctive feature of AI-Pastiche, central to this study, is the integration of metadata with synthetic human evaluations of generated samples along three key dimensions: the perceived authenticity of the artwork, its adherence to the prompt, and a defect score reflecting the presence and severity of visible artifacts. These synthetic scores were derived by aggregating human ratings collected through large-scale user surveys.

Pronounced anatomical distortions—especially in the rendering of hands and facial features—remain common in outputs produced by generative models. Representative examples are shown in Figure 5. The ability to accurately recognize such defects is a fundamental prerequisite for artistic evaluation. A central goal of this study is to assess whether CLIP-based embeddings are sufficiently expressive to support their detection.

In summary, the AI-pastiche dataset enables a direct comparison between human and machine vision models—precisely the objective of this investigation.





Fig. 5. Examples from the AI-Pastiche dataset showcasing major structural and anatomical artifacts.

## 4 Methodology

Our aim is to investigate CLIP's ability to extract high-level semantic and stylistic information from paintings and to evaluate its perceptual capabilities across multiple dimensions, including content and scene understanding, artistic style, correlation with human perception, and detection of visual deformations or artifacts.

In the following sections, we shall detail the experiments conducted in order to assess the previous aspects of CLIP's vision mechanism. All experiments are conducted without fine-tuning or architectural modification. This design choice is deliberate: fine-tuning would adapt the model to the task, whereas our objective is to analyze the perceptual priors and representational structure learned during CLIP's original pretraining.

### 4.1 Image-Text Alignment

This experiment evaluated CLIP's ability to align visual representations of artworks with their corresponding textual descriptions or prompts, that its core cross-modal capability. The task establishes a baseline for understanding how effectively CLIP's shared embedding space captures semantic relationships between artistic images and language.

For each image ( $I_i$ ) and text snippet ( $T_j$ ), CLIP's image and text encoders project them into a common feature space as  $(v_i = f_I(I_i))$  and  $(t_j = f_T(T_j))$ . Their semantic proximity is then quantified using cosine similarity:  $S_{i,j} = \frac{v_i \cdot t_j}{|v_i| |t_j|}$  which provides a direct measure of the alignment between the visual and linguistic modalities.

The experiment was conducted across the three datasets that differ in origin, structure, and degree of semantic variability. The NGAD comprises 1,521 curated artworks paired with detailed museum descriptions. Here, the task was formulated as an information retrieval problem: each image was compared against all available summaries, and performance was evaluated using the Recall@k metric, measuring how often the correct text appeared among the top  $k$  most similar candidates.

The AI-Pastiche Dataset, by contrast, contains images synthetically generated from a fixed set of 73 prompts. Because of this limited and repetitive vocabulary, the task was treated as a multi-class classification problem, where accuracy represented the proportion of images correctly matched to their original prompts.

Finally, for AI-WikiArt Dataset, the experiment focused on the subset of 118,590 AI-generated images, produced from 39,471 unique prompts—each rendered by three different generative models. For each image, alignment was computed against its corresponding prompt, framing the task as information retrieval and evaluating performance using the Recall@k metric.

A key limitation encountered was CLIP's maximum text input length of 77 tokens, whereas the descriptions or the prompt in the case of the NGA dataset often exceeded this constraint. To address this, we generated concise summaries of each description using ChatGPT 4o-mini, explicitly including the painting's subject, style, and

period as input to ensure that these elements were preserved in the summaries. The exact prompt used for this summarization process is reported in Appendix A.

All experiments were conducted across multiple CLIP model variants available in the OpenAI library [37], including both ResNet- and **Vision Transformer (ViT)**-based architectures, in order to compare performance across different model capacities and input resolutions.

## 4.2 Style Recognition

In this second experiment, we investigate CLIP’s capability to correctly associate artworks with their corresponding artistic styles.

Our first experiment was a simple image to sentence correlation, comparing in CLIP’s embedding space an image to a standardized textual prompt of the form “an artwork in [style] style.” Using CLIP’s image and text encoders, we computed L2-normalized embeddings for both modalities. The cosine similarity between each image embedding and all style prompts was then calculated, producing a similarity matrix that quantifies the degree of semantic alignment between images and candidate styles.

The performance was assessed using Recall@ $k$  metrics, where Recall@ $k$  represents the proportion of test images for which the correct style is ranked within the top  $k$  most similar textual prompts.

As we shall see, and as it was also partially expected, this technique does not work well for style recognition, to the strong modality gap between images and text.

To address these limitations, our second experiment consisted of training a linear classifier from the image embeddings, a technique sometimes referred to as linear probe.

We consider a dataset  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ , where each image  $x_i$  is annotated with a style label  $y_i \in \{1, \dots, C\}$ . Two sets of embeddings are extracted:

- Visual embeddings  $v_i \in \mathbb{R}^D$  are obtained by encoding each image with CLIP’s image encoder.
- Textual embeddings  $t_c \in \mathbb{R}^D$  are computed for each style  $c$  by encoding the prompt “an artwork in style” with CLIP’s text encoder.

Both visual and textual embeddings are L2-normalized.

To investigate the structure of each modality’s embedding space, we independently train multiclass logistic regression classifiers on the visual and textual embeddings.

For the visual modality, the prediction for image  $i$  is:

$$\hat{y}_i = \arg \max_c w_c^{\text{img}\top} v_i,$$

where  $w_c^{\text{img}}$  denotes the weight vector associated with style  $c$ , forming the matrix  $W_{\text{img}} \in \mathbb{R}^{C \times D}$ . The accuracy of this prediction gives a first interesting result about the possibility to exploit CLIP’s embedding for style recognition.

In addition, we can train a similar classifier for textual embeddings and then use its inversion to project back the classification into CLIP’s embedding space, obtaining, by composition, a linear transformation what should bring each embedding close to the corresponding style sentence.

Specifically, suppose that the classifier trained on textual embeddings yields a weight matrix  $W_{\text{txt}} \in \mathbb{R}^{C \times D}$  and predictions:

$$\hat{y}_c = \arg \max_{c'} w_{c'}^{\text{txt}\top} t_c.$$

The linear transformation

$$T = W_{\text{txt}}^\top W_{\text{img}} \in \mathbb{R}^{C \times C}$$

can be applied to a visual embedding  $v_i$ , obtaining

$$v'_i = Tv_i,$$

which is intended to be semantically aligned with its textual counterpart. The similarity between a transformed visual embedding  $v'_i$  and a textual style embedding  $t_c$  is computed as:

$$S_{c,i} = t_c^\top v'_i = t_c^\top T v_i.$$

This produces a similarity matrix  $S \in \mathbb{R}^{C \times N}$ , where each column corresponds to the similarity scores between a given image and all style descriptions.

### 4.3 Adherence Analysis

While the previous experiments focused on classification and retrieval tasks, this experiment addresses a more subjective dimension, namely whether CLIP's similarity scores reflect human judgments of how well images derived from a same prompt respect the requirements. Different from the first class of experiment, this a fine-grained intraclass evaluation among images relative to a common prompt.

In [4], the authors conducted an adherence survey in which participants were presented with multiple images generated by different models from the same textual prompt. Since contemporary generative models typically produce images that adhere reasonably well to their prompts, an absolute evaluation of prompt alignment would have limited discriminative value. Instead, participants were asked to provide a comparative judgment, rating each image as good, neutral, or bad relative to the other images generated from the same prompt. By aggregating responses across participants, the authors derived a prompt-specific adherence score for each image, reflecting its perceived stylistic and semantic alignment in comparison to alternative realizations of the same textual description. This average adherence aligns sufficiently well with each individual human evaluation: the average cosine similarity is around 0.7, with a standard deviation of 0.06.

To explore whether CLIP captures similar perceptual cues, we conducted a parallel analysis using CLIP similarity scores. For each prompt, we computed the cosine similarity between the CLIP text embedding of the prompt and the CLIP image embeddings of all associated generated images. These scores were then normalized and scaled to produce a vector directly comparable to the human-derived adherence scores.

Finally, we computed the cosine similarity between the CLIP similarity vector and the human adherence scores, using it as a synthetic measure of alignment between CLIP's assessments and human perception. This procedure was repeated across all available CLIP models.

### 4.4 Perception of Artifacts and Deformations

The final experiment examined whether CLIP's visual representations are sensitive to the technical quality of AI-generated images, particularly the presence of visual artifacts such as distortions, inconsistencies, and anatomical errors. While humans often judge prompt adherence in part through these cues, CLIP may overlook them due to its focus on high-level semantic content.

According to the instruction provided in [4], the user was asked to do an adherence evaluation performed by humans seems to be essentially driven by three different directions:

- Content: respect for the subject described in the prompt;
- Style: respect for the style that the generative model was required to imitate;
- Overall Quality: absence of visible artifacts and deformations typical of generative techniques.

One of the surveys presented in [4] specifically targeted the presence of visible artifacts or deformations in generated images. Defects were categorized as:

- *Major*: clearly visible or frequent errors (e.g., severe anatomical mistakes);

- *Minor*: less critical imperfections (e.g., extra fingers, small distortions);
- *None*: no apparent visual defects.

Each image was assigned a *defect score* averaging these evaluations over a large number of participants.

In the first stage, we tested whether information about artifacts could be linearly decoded from CLIP’s image embeddings by regressing the human defect scores against the corresponding CLIP features. The near-zero  $R^2$  values indicated that CLIP’s global representations do not explicitly encode artifact-related information.

In the second stage, to explore whether defect information could improve CLIP’s prompt adherence estimation, we introduced a linear combination model that integrates CLIP-based perceptual similarity scores with defect scores.

Given the normalized perceptual similarity  $\hat{s}_i$  for image  $i$  and its defect score  $d_i$ , the predicted quality score  $\tilde{y}_i$  is computed as:

$$\tilde{y}_i = a \hat{s}_i + b d_i + c, \quad (1)$$

where  $a$  and  $b$  are learned coefficients and  $c$  is a bias term. The coefficients are estimated via the least squares method:

$$\min_{a,b,c} \sum_{i=1}^n (y_i - (a \hat{s}_i + b d_i + c))^2 \quad (2)$$

with  $y_i$  denoting the human-annotated adherence score. Finally, the amended similarity vector  $\tilde{\mathbf{y}}$  is compared to the human adherence vector  $\mathbf{y}$  using cosine similarity:

$$\text{sim}(\tilde{\mathbf{y}}, \mathbf{y}) = \frac{\tilde{\mathbf{y}} \cdot \mathbf{y}}{\|\tilde{\mathbf{y}}\| \|\mathbf{y}\|}. \quad (3)$$

## 5 Results on Image–Text Alignment

This first experiment assessed CLIP’s ability to align artworks with their corresponding textual descriptions across both human-created and AI-generated datasets. The results consistently confirm that CLIP maintains strong visual–semantic correspondence within the artistic domain, although performance varies depending on model scale, input modality, and the linguistic characteristics of the accompanying text.

On the NGAD, the task was formulated as image-to-summary retrieval. As shown in Table 1, all CLIP variants achieved high Recall@k scores, with a clear performance increase correlated with model size and input resolution. Among the ResNet-based architectures, performance rose steadily from the baseline RN50 (Recall@1)=66.3 to RN50x64 (Recall@1)=82.8 for distinguishing between artworks with similar subjects or styles.

A comparable trend emerged among the ViT variants, with higher-capacity models and higher image resolutions yielding superior results. The ViT-L/14@336px model reached the best overall performance. The trend is also confirmed with open CLIP models. The very large PE-Core-bigG-14, with a latent space of 1,280 features, is the model showing best performances on this task, achieving an impressive Recall@1=93.4% and Recall@5=99.1%.

To assess the performance on synthetic imagery, the same alignment procedure was applied to the AI-Pastiche dataset, which contains multiple generated images for 73 unique prompts. Treated as a multi-class classification task, all models achieved high accuracy ( $\geq 86\%$ ), with RN50x64 and ViT-L/14 variants reaching 89.6%, while EVA02-E-14-plus and PE-Core-bigG-14 achieved around 94%. The results for the AI-Pastiche dataset are reported in Table 2 for characteristic models.

We only report accuracy in this case, due to the limited number of prompts. From a generative standpoint, the high accuracy values indicate that the image generation process produced outputs closely aligned with the subjects described in the prompts.



Table 1. Summary–Image Alignment for NGAD Images  
(Image→Text)

| Model           | Recall@1     | Recall@5     | Recall@10    |
|-----------------|--------------|--------------|--------------|
| RN50            | 0.663        | 0.915        | 0.966        |
| RN101           | 0.693        | 0.926        | 0.966        |
| RN50x4          | 0.741        | 0.946        | 0.978        |
| RN50x16         | 0.791        | 0.964        | 0.988        |
| RN50x64         | 0.828        | 0.970        | 0.990        |
| ViT-B/32        | 0.678        | 0.925        | 0.970        |
| ViT-B/16        | 0.709        | 0.928        | 0.969        |
| ViT-L/14        | 0.794        | 0.972        | 0.989        |
| ViT-L/14@336px  | 0.814        | 0.974        | 0.991        |
| ViT-H-14-378    | 0.859        | 0.984        | 0.993        |
| EVA02-E-14-plus | 0.907        | 0.991        | 0.998        |
| ViT-SO400M-14   | 0.850        | 0.968        | 0.986        |
| ViT-SO400M-16   | 0.834        | 0.971        | 0.992        |
| CoCa-L-14       | 0.841        | 0.970        | 0.987        |
| ConvNeXt-XXL    | 0.890        | 0.988        | 0.997        |
| ViT-bigG-14     | 0.743        | 0.935        | 0.971        |
| MobileCLIP2-L   | 0.806        | 0.962        | 0.983        |
| ViTamin-XL      | 0.866        | 0.982        | 0.994        |
| PE-Core-bigG-14 | <b>0.934</b> | <b>0.991</b> | <b>0.997</b> |
| NLLB-CLIP-large | 0.605        | 0.854        | 0.920        |

The first set refers to OpenAI models, while the second set are open models.  
Best results in bold.

Table 2. Accuracy of Different CLIP Models in Matching Generated Images  
with Their Corresponding Summarized Prompts in the AI-Pastiche Dataset

| Model           | ViT-L/14@336px | EVA02-E-14-plus | PE-Core-bigG-14 |
|-----------------|----------------|-----------------|-----------------|
| <b>Accuracy</b> | 0.896          | 0.938           | 0.941           |

Table 3. Image–Prompt Alignment Performance for the AI-Generated Subset of the WikiArt Dataset, Measured for  
ViT-L/14@336px, EVA02-E-14-plus, and PE-Core-bigG-14

|                        | ViT-L/14@336px |        |        | EVA02-E-14-plus |        |        | PE-Core-bigG-14 |        |        |
|------------------------|----------------|--------|--------|-----------------|--------|--------|-----------------|--------|--------|
|                        | Rec@1          | Rec@5  | Rec@10 | Rec@1           | Rec@5  | Rec@10 | Rec@1           | Rec@5  | Rec@10 |
| WikiArt (AI-generated) | 0.4258         | 0.6565 | 0.7416 | 0.5568          | 0.7662 | 0.8341 | 0.6504          | 0.8400 | 0.8928 |

Larger models provide a better alignment.

The AI-WikiArt dataset posed a considerably more challenging retrieval task. The most interesting experiment is relative to the AI-generated subset, comprising 118,590 images produced from 39,471 unique prompts, each rendered by three different generative models. The results for the AI-WikiArt dataset are reported in Table 3.

Performance was notably lower than on NGAD, primarily due to the larger retrieval space (nearly 40,000 candidate prompts), the greater linguistic variability of LLM-generated descriptions, and the modest achievements of generators. Furthermore, the low performance on AI-WikiArt is also a result of the prompts themselves not

**True Prompt** (Similarity = 0.249):

The image depicts a serene landscape with a tall tree, a small house, and a forested background, created using an impressionistic technique with soft, blended brushstrokes and muted colors.

**Predicted Prompt** (Similarity = 0.293):

The image depicts a landscape with trees and a rocky structure, created using the Post-Impressionist technique of layered, geometric brushstrokes characteristic of Paul Cézanne's style.

Generator: Stable-Diffusion

Style: Cubism

**True Prompt** (Similarity = 0.320):

The image depicts a vibrant tropical garden with palm trees and a rustic cottage, created using impressionist brushstrokes and a warm color palette.

**Predicted Prompt** (Similarity = 0.328):

The image depicts a sunlit garden scene with a large palm tree, lush greenery, a small house with a red roof, and a figure in the background, created using soft, impressionistic brushstrokes.

Generator: F-Lite

Style: Impressionism

Fig. 6. Examples of prompt–image alignment failures on the AI-WikiArt dataset for the ViT-L/14@336px model. The images were incorrectly matched, but the retrieved prompts are semantically very similar to the ground-truth prompts, explaining the low retrieval scores.

being well curated and often being highly semantically similar to each other. This lack of distinctiveness makes the retrieval task exceptionally challenging, as a given image may be plausible for many similar prompts, as shown in the examples in Figure 6.

The examples in Figure 6 highlight this problem perfectly. In both cases, the model failed to retrieve the correct prompt. However, the Predicted Prompt (the top-ranked incorrect choice) not only achieved a higher similarity score than the True Prompt, but it is also semantically almost identical. In the first example, both prompts describe a landscape with trees and a building, and the model's top pick is arguably a better description of the Cézanne-like image, even though it is not the ground truth. In the second example, both prompts describe a “garden scene with ... palm trees and a ... house” using “impressionistic brushstrokes.” This demonstrates that the prompts are not distinct enough for a meaningful retrieval task, leading to low quantitative scores that reflect the dataset's ambiguity rather than a complete failure of the model's semantic alignment.

## 6 Results on Style Recognition

Following our analysis of content-based classification, we now assess CLIP's more abstract capability for artistic style recognition. This task tests the model's sensitivity to subtle visual features like brushwork, color palette, and composition, which is a known challenge for models biased toward content over form.

To quantitatively evaluate CLIP's competence in this task, we conducted a multi-stage analysis. Following preliminary experiments, we focused on three CLIP variants: OpenAI's ViT-L/14@336px, EVA02-E-14-plus, and PE-Core-bigG-14. These models differ in latent embedding dimensionality (768, 1,024, and 1,280, respectively),

Table 4. Results of Artistic Style Recognition on Different Datasets Using a Zero-Shot Approach, Based on the Similarity between an Image and Its Style Label

|                             | ViT-L/14@336px |        |        | EVA02-E-14-plus |        |        | PE-Core-bigG-14 |        |        |
|-----------------------------|----------------|--------|--------|-----------------|--------|--------|-----------------|--------|--------|
|                             | Acc.           | F1     | Rec@3  | Acc.            | F1     | Rec@3  | Acc.            | F1     | Rec@3  |
| <b>NGAD</b>                 | 0.3006         | 0.2303 | 0.6080 | 0.5432          | 0.4547 | 0.5150 | 0.5556          | 0.3907 | 0.7860 |
| <b>AI-Pastiche</b>          | 0.4974         | 0.3830 | 0.7696 | 0.5079          | 0.4051 | 0.7749 | 0.4764          | 0.4096 | 0.7906 |
| <b>AI-WikiArt (H + AI)</b>  | 0.2503         | 0.2425 | 0.4906 | 0.2483          | 0.2028 | 0.5136 | 0.2683          | 0.2520 | 0.5531 |
| <b>AI-WikiArt (Human)</b>   | 0.4087         | 0.3642 | 0.6690 | 0.3759          | 0.4725 | 0.6493 | 0.4492          | 0.3757 | 0.7350 |
| <b>AI-WikiArt (AI)</b>      | 0.1956         | 0.1879 | 0.4288 | 0.1548          | 0.1759 | 0.3828 | 0.2066          | 0.1938 | 0.4891 |
| <b>AI-ArtBench (H + AI)</b> | 0.6378         | 0.6367 | 0.8830 | 0.6541          | 0.6923 | 0.8985 | 0.6483          | 0.6402 | 0.9147 |
| <b>AI-ArtBench (Human)</b>  | 0.5516         | 0.5156 | 0.8336 | 0.5434          | 0.5009 | 0.8257 | 0.5880          | 0.5713 | 0.8800 |
| <b>AI-ArtBench (AI)</b>     | 0.6754         | 0.6841 | 0.9053 | 0.7066          | 0.6798 | 0.9350 | 0.6841          | 0.6775 | 0.9313 |

The evaluated models are ViT-L/14@336px, EVA02-E-14-plus, and PE-Core-bigG-14. All models show modest performance, reflecting the fact that artistic style is not a dominant component in CLIP embeddings, as somewhat expected.

allowing us to assess whether representational capacity influences stylistic discrimination. Additional results for other CLIP architectures are reported in Appendix C.

For the selected models, we first established a baseline using zero-shot classification with style-based text prompts. We then examined whether stylistic information is linearly separable in the visual embedding space through supervised methods, including a linear probe (logistic regression) applied to frozen image embeddings and a cross-modal alignment approach.

### 6.1 Baseline Zero-Shot CLIP Performance on Style Classification

We first established a baseline for CLIP’s intrinsic style recognition capabilities using a zero-shot classification approach. Image embeddings were compared against text embeddings generated from prompts formatted as “an artwork in [style] style.” The results are shown in Table 4.

The differing trends observed across datasets deserve brief clarification. In AI-WikiArt, zero-shot performance is higher for human-created artworks than for AI-generated samples, whereas AI-ArtBench exhibits the opposite pattern. These differences likely reflect structural characteristics of the datasets rather than genuine variations in CLIP’s stylistic sensitivity. AI-ArtBench contains a limited number of well-separated style classes and relatively controlled generation procedures, which may reduce intra-class variability and facilitate discrimination in a zero-shot setting. In contrast, AI-WikiArt encompasses a broader stylistic and authorial diversity, greater semantic variability, and more heterogeneous generative outputs, introducing distributional complexity that may weaken direct text–image alignment.

Importantly, despite these dataset-specific effects, the overarching result remains consistent: zero-shot style recognition performance is modest and unstable across conditions, reinforcing the conclusion that stylistic information is not robustly aligned with textual prompts in CLIP’s native embedding space.

To better investigate the modal gap highlighted in the previous results, we employed the UMAP algorithm [32] to generate a 3D projection of the high-dimensional embeddings. These embeddings included both image representations and textual prompts describing artistic styles (e.g., “an artwork in [style] style”).

We mainly conducted this investigation for the lighter ViT-L/14@336px model. The resulting visualization (Figure 7) reveals a clear separation between textual encodings (the small cluster on the left) and image encodings (the large cluster on the right).

For images of paintings, we also distinguish between correctly classified samples, shown as green bullets, and misclassified ones, shown as red crosses. The substantial entanglement of these two classes suggests a dominance

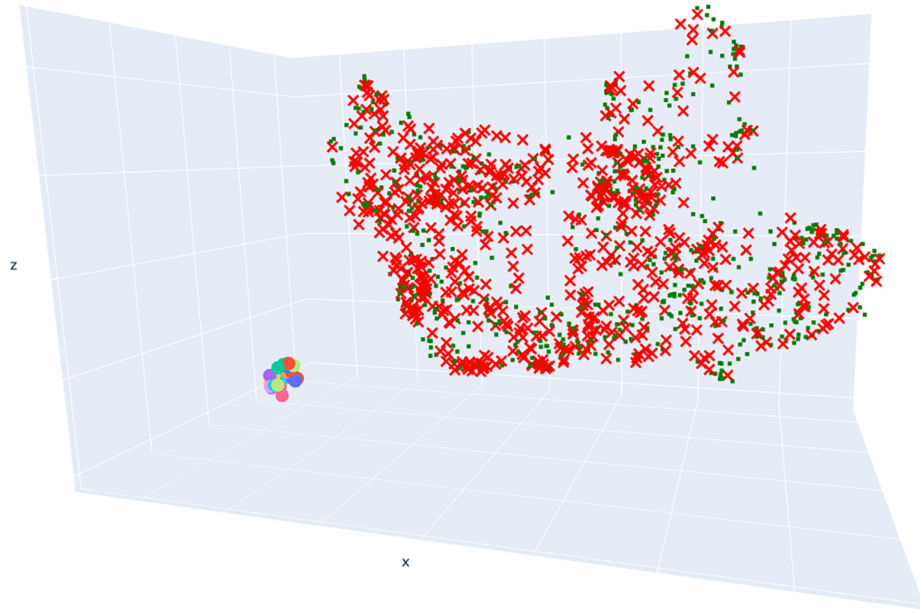


Fig. 7. 3D projection of the textual embeddings of artistic styles and the visual embeddings of correctly classified (green) and misclassified (red) images using UMAP.

of non-stylistic features in the embeddings. However, the chaotic pattern could also be a consequence of the aggressive dimensionality reduction and may not accurately reflect the semantic structure present in the original high-dimensional space.

To further clarify the issue, we conducted a detailed analysis based on nearest neighbors. Specifically, we focused on close-up image pairs in which one image was correctly classified, while the other was not. By identifying such pairs, we were able to isolate semantically coherent examples in which classification performance diverged. This setup provided valuable insight into how subject matter similarity and stylistic cues interact within CLIP’s representation space, and where the model may struggle to disentangle the two.

Representative examples are shown in Figure 8, where the misclassified image is displayed on the left, and its nearest correctly classified neighbor appears on the right. For each image, we report the cosine similarity scores with respect to both the true and the predicted style prompts, computed in both the original latent space and its lower-dimensional projection. This comparison allows us to examine how proximity in the embedding space relates to classification outcomes and whether stylistic distinctions are preserved through dimensionality reduction.

These findings underscore a key limitation of CLIP: it tends to encode and prioritize semantic content (such as objects, scenes, and compositions) over stylistic features such as brushwork, color palette, or compositional structure. This leads to frequent misclassifications when artworks differ in style but share similar subject matter. This bias is consistent with CLIP’s training objectives, which favor semantic alignment based on large-scale image–text data, where captions often emphasize content over formal style. As a result, CLIP’s latent space tends to conflate stylistically diverse images with similar semantics.

Despite these limitations, in low-dimensional projections of embeddings of NAG paintings, there is an apparent organization of data in style-related clusters (see Appendix C).

In the next sections, we shall explore supervised techniques projecting the embedding space on semantically instructed subspaces, following the methodology described in Section 4.2.



Actual style: Gothic - Similarity: 0.151  
Predicted style: Renaissance - Similarity: 0.196



Actual style: Renaissance - Similarity: 0.203  
Predicted style: Renaissance - Similarity: 0.203



Actual style: Post-Impressionist - Similarity: 0.167  
Predicted style: Academic - Similarity: 0.171



Actual style: Post-Impressionist - Similarity: 0.182  
Predicted style: Post-Impressionist - Similarity: 0.182



Actual style: Realist - Similarity: 0.173  
Predicted style: Baroque - Similarity: 0.200



Actual style: Baroque - Similarity: 0.223  
Predicted style: Baroque - Similarity: 0.223



Fig. 8. Visual comparison between a misclassified image (left) and a correctly classified image (right). For each artwork, the actual and predicted artistic styles are shown, along with their respective similarity scores to the image in the latent space.



Table 5. Intra-Dataset Results of Artistic Style Recognition with Linear Probe Method Using Unnormalized Embeddings

|                            | ViT-L/14@336px |        | EVA02-E-14-plus |        | PE-Core-bigG-14 |        |
|----------------------------|----------------|--------|-----------------|--------|-----------------|--------|
|                            | Accuracy       | F1     | Accuracy        | F1     | Accuracy        | F1     |
| <b>NGAD</b>                | 0.7771         | 0.7775 | 0.8683          | 0.8689 | 0.8930          | 0.8909 |
| <b>AI-WikiArt (Human)</b>  | 0.7528         | 0.7543 | 0.7808          | 0.7818 | 0.7822          | 0.7865 |
| <b>AI-ArtBench (Human)</b> | 0.7495         | 0.7491 | 0.7744          | 0.7742 | 0.7953          | 0.8015 |

Table 6. Cross-Dataset Generalization Results with a Linear Probe Trained on AI-WikiArt (Human)

|                                      | ViT-L/14@336px |        | EVA02-E-14-plus |        | PE-Core-bigG-14 |        |
|--------------------------------------|----------------|--------|-----------------|--------|-----------------|--------|
|                                      | Accuracy       | F1     | Accuracy        | F1     | Accuracy        | F1     |
| <b>AI-WikiArt Human (test split)</b> | 0.7654         | 0.7670 | 0.7904          | 0.7920 | 0.8070          | 0.8081 |
| <b>NGAD</b>                          | 0.4729         | 0.4831 | 0.7200          | 0.7363 | 0.6696          | 0.7048 |
| <b>AI-Pastiche</b>                   | 0.3452         | 0.3682 | 0.3559          | 0.3938 | 0.3511          | 0.3773 |
| <b>AI-ArtBench (Human+AI)</b>        | 0.5267         | 0.5498 | 0.5190          | 0.5589 | 0.4404          | 0.5114 |
| <b>AI-ArtBench (Human)</b>           | 0.6024         | 0.6068 | 0.6244          | 0.6448 | 0.5562          | 0.6061 |
| <b>AI-ArtBench (AI)</b>              | 0.4905         | 0.5093 | 0.4686          | 0.5091 | 0.3875          | 0.4627 |
| <b>AI-WikiArt (AI)</b>               | 0.3088         | 0.3087 | 0.3302          | 0.3171 | 0.3037          | 0.3193 |

## 6.2 Style Classification Performance with Supervised Methods

*Intra-Dataset Linear Probing.* To evaluate the representational power of the visual encoder, we first conducted a linear probing experiment by training a linear classifier (logistic regression) on an 80% split of each dataset and testing on the remaining 20%. An interesting preliminary finding was that the use of unnormalized embeddings sensibly improved accuracy (e.g., from 70.74% to 77.71% on NGAD, with ViT-L/14@336px), suggesting that vector magnitude carries meaningful stylistic information. The results with unnormalized embeddings are shown in Table 5.

We only report the results for Human subsets of the dataset since:

- training on generated images takes into account the implicit biases of the generators (that is, the style as understood by the generators). Essentially, we are evaluating the internal coherence of generated images.
- due to the limited number of prompts, the accuracy for AI-Pastiche and especially of AI- would be inflated as the random 80/20 split of the small dataset likely led to near-duplicate images from identical prompts appearing in both train and test sets.

The accuracy of ViT-L/14@336px is similar for all datasets, around 75–78%. Both EVA02-E-14-plus and PE-Core-bigG-14 show excellent performance on NGAD. However, their performance is not matched on different datasets, hinting at possible contamination during training.

*6.2.1 Cross-Dataset Generalization Test.* To test the robustness of the classification, we performed an additional cross-dataset evaluation, training a logistic regression classifier exclusively on the AI-WikiArt (Human) training data and testing it on all other datasets. In this case, it is interesting to test the classifier even on generated images, to check their stylistic quality. The results, shown in Table 6, seem to show a strong domain gap. Performance dropped significantly across all datasets, revealing a lack of understanding of abstract and generalizable style concepts.

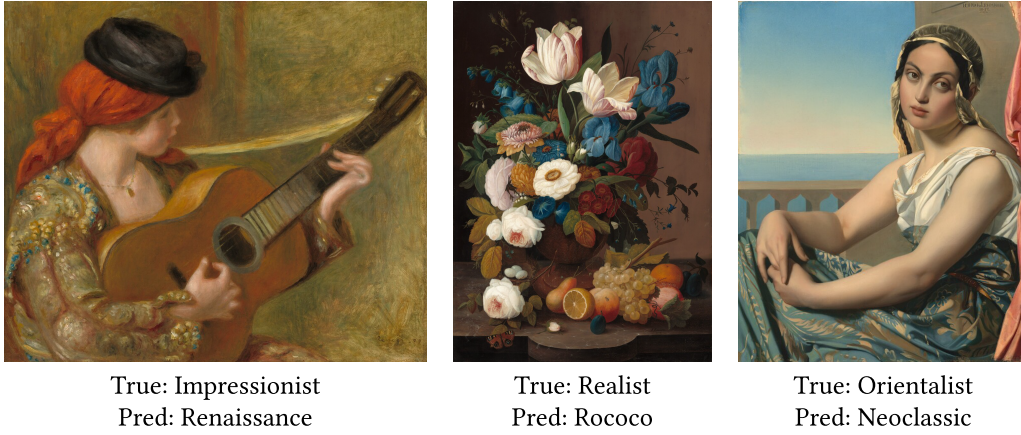


Fig. 9. Examples of misclassifications in style recognition.

Table 7. Intra-Dataset Results of Artistic Style Recognition with Supervised Cross-Modal Alignment Method

|                          | ViT-L/14@336px |        |        | EVA02-E-14-plus |        |        | PE-Core-bigG-14 |        |        |
|--------------------------|----------------|--------|--------|-----------------|--------|--------|-----------------|--------|--------|
|                          | Acc.           | F1     | Rec@3  | Acc.            | F1     | Rec@3  | Acc.            | F1     | Rec@3  |
| <b>NGAD</b>              | 0.7429         | 0.6304 | 0.9554 | 0.8272          | 0.8221 | 0.9877 | 0.7901          | 0.7364 | 0.9835 |
| <b>AI-WikiArt Human</b>  | 0.5931         | 0.5554 | 0.8796 | 0.6353          | 0.5763 | 0.8886 | 0.5511          | 0.4919 | 0.8595 |
| <b>AI-ArtBench Human</b> | 0.6809         | 0.6763 | 0.9170 | 0.7001          | 0.7003 | 0.9359 | 0.5296          | 0.4581 | 0.8354 |

Some examples of misclassifications relative to the NGA dataset and the ViT-L/14@336px model are given in Figure 9. Each image is shown with its true style label and the incorrect prediction made by the best-performing model.

Taking into account generated images, the modest results clearly stress the many deficiencies of current generators in replicating artistic styles, as also pointed out in [4]. The slightly better, yet still mediocre, performance on AI-ArtBench AI (49.05%) is likely to be explained in terms of a “shared inductive bias.” Generative models used for AI-ArtBench, such as Latent Diffusion and Standard Diffusion, often use CLIP as a core component in their generation process. These models were tasked with a simple prompt (“A painting in <art style> art style”), with variation introduced only by changing the seed. As a result, they produce outputs that are optimized for CLIP’s perceptual space, causing images of the same intended style to be visually very similar and to cluster tightly in the embedding space, as already shown in Section 2.

**6.2.2 Performance with Supervised Cross-Modal Alignment.** Finally, we investigated if learning a linear transformation ( $T$ ) to map visual embeddings to the text space could bridge the “modality gap.” The results on intra-dataset splits (Table 7) showed substantial improvements over the zero-shot baseline. This shows that the modality gap is structurally coherent within a given dataset.

**6.2.3 Benchmark against Few-Shot Adaptation (APE).** To contextualize our findings, we benchmarked our methods against **Adaptive Prior Refinement (APE)**, a state-of-the-art few-shot learning technique. We evaluated APE on the NGAD dataset using a varying number of training examples (“shots”) per class to simulate data-scarce scenarios. We followed an 80/10/10 split into training, validation, and test sets. Few-shot settings were simulated by randomly sampling a fixed number of examples per class. APE’s feature selection process reduced the original 768-dimensional CLIP “ViT-L/14@336px” embeddings to 384 refined features. APE’s performance scaled with the number of shots, with accuracy increasing from 37.21% (1-shot) to a peak of 53.42% (16-shot), as shown in Table 8.

Table 8. APE Model Accuracy and Recall@K Computed in NGAD Test Split

| Shots | Accuracy | Recall@1 | Recall@3 | Recall@5 |
|-------|----------|----------|----------|----------|
| 1     | 37.21    | 37.21    | 64.16    | 78.31    |
| 2     | 38.13    | 38.13    | 73.29    | 84.93    |
| 4     | 50.23    | 50.23    | 79.00    | 90.18    |
| 8     | 41.10    | 41.10    | 74.43    | 88.36    |
| 16    | 53.42    | 53.42    | 84.93    | 91.55    |

Even in its optimal 16-shot configuration, APE’s performance remained significantly below that of our supervised methods trained on the full dataset. Both cross-modal alignment (74.29% accuracy) and linear probe (77.71% accuracy) using unnormalized embeddings proved considerably more effective.

This comparison is instructive within the evaluated setting. For the NGAD dataset, where a sufficient amount of labeled data is available, a supervised linear classifier trained on the full embedding space achieved higher accuracy than the tested few-shot adaptation method (APE). This result suggests that stylistic information in CLIP’s embeddings may be broadly distributed across dimensions, rather than confined to a limited subset of features selected by few-shot refinement strategies.

We emphasize, however, that this observation is restricted to the specific adaptation technique and experimental configuration considered here. A broader comparison with additional adaptation frameworks would be necessary to draw general conclusions. Within the scope of our study, the results indicate that, when sufficient supervision is available, a simple linear probe can effectively exploit the information already present in the embedding space.

## 7 Correlation with Human Judgment

To assess the alignment between CLIP’s internal metrics and qualitative human perception, we compared the model’s scores against human evaluations from the AI-Pastiche dataset. These experiments aim to quantify the perceptual gap between machine-based and human judgments. Specifically, in the following sections, we focus on adherence to generative prompts (Section 7.1) and the perception of artifacts and deformations (Section 7.2).

A related but more delicate dimension concerns authenticity, namely the distinction between human-created and AI-generated artworks. As observed in recent studies [26], CLIP-based representations can be highly effective in discriminating between real and synthetic images, often achieving performance levels that exceed those of human evaluators in controlled settings. However, the reasons underlying this capability remain largely unclear. In light of the other results presented in this article, it seems unlikely that such performance reflects a genuine sensitivity to stylistic or compositional cues; rather, it is more plausibly explained by the exploitation of subtle statistical regularities and artifacts introduced by generative models. Interpreting this behavior is particularly challenging, as it raises questions about the nature of the features encoded in CLIP’s latent space and their relation to human notions of authenticity. We view this as an important and delicate open problem, which we are currently investigating in ongoing work.

### 7.1 Adherence to Generative Prompts

First, we measured the correlation between CLIP’s prompt-image similarity scores and human ratings of prompt adherence. We found a moderate positive correlation, with the model that performed the best (ViT-L/14@336px), achieving a cosine similarity of 0.437 with human judgment (Table 9).

This value, while indicating some alignment, is substantially lower than the average human inter-annotator correlation of approximately 0.7 (see Section 4.3). This quantitatively confirms a significant divergence between the criteria used by AI and human observers to judge prompt adherence.

Table 9. Cosine Similarity between the Vector of CLIP's Adherence Scores and the Vector of Average Human Judgments

| Model               | Alignment with Human Judgment | After Defect Score Integration |
|---------------------|-------------------------------|--------------------------------|
| RN50                | 0.406                         | 0.478                          |
| RN101               | 0.398                         | 0.462                          |
| RN50x4              | 0.379                         | 0.448                          |
| RN50x16             | 0.413                         | 0.489                          |
| RN50x64             | 0.428                         | 0.484                          |
| ViT-B/32            | 0.411                         | 0.481                          |
| ViT-B/16            | 0.383                         | 0.458                          |
| ViT-L/14            | 0.425                         | 0.482                          |
| ViT-L/14@336px      | <b>0.437</b>                  | 0.497                          |
| EVA02-E-14-plus     | 0.434                         | <b>0.501</b>                   |
| PE-Core-bigG-14-448 | 0.419                         | 0.490                          |

The second column shows the improved alignment when human-annotated defect scores are integrated into the model. All values should be compared with the average human similarity to the adherence score, that is around  $0.7 \pm 0.06$ . Highest values in bold.

## 7.2 Perception of Artifacts and Deformations

Next, we investigated CLIP's sensitivity to visual artifacts and technical flaws, a key factor in human aesthetic judgment. An initial attempt to predict human-annotated defect scores directly from CLIP's image embeddings failed completely, yielding a coefficient of determination ( $R^2$ ) near zero. This provides strong evidence that the model is effectively "blind" to low-level structural flaws.

We then tested whether this blindness contributes to the perceptual gap. By integrating human-annotated defect scores into a simple model along with CLIP's similarity score (described in Section 4.4), the correlation with human adherence ratings improved consistently for all models. For the best-performing model, the alignment score increased from 0.434 to 0.501 (to be compared with the average human alignment to the adherence score, which is around 0.7). Although an improvement of 0.06 may appear modest, it is of the same order of magnitude as the standard deviation of human evaluations, and we therefore consider it a significant indication of the lack of features related to defects and artifacts in CLIP embeddings.

Despite this improvement, CLIP occasionally assigns high adherence scores to images rated poorly by humans, and vice versa. Such mismatches are often attributable to divergences in content or style rather than the presence of artifacts. This demonstrates that a substantial portion of the discrepancy between CLIP and human judgment is due to the model's inability to perceive technical quality.

To visualize these results, a qualitative analysis of the discrepancies was also performed. CLIP's evaluation is heavily biased toward semantic content, often overlooking compositional, stylistic, or relational errors that are immediately apparent to humans. Some typical examples are given in Figure 10.

Figure 10(a) was supposed to represent a fight between two knights on horseback; humans penalized the fact that only one knight is represented, while CLIP does not seem to notice the absence. Figure 10(b), generated by Auto-Aesthetics V1, was meant to depict a lively Parisian street scene in the rain, in a style resembling impressionism: humans reproach the lack of adherence to the required style. The prompt for Figure 10(c) required the representation of a kneeling figure in rich robes offering a white flower to another figure, draped in blue; the fact that one of the two figures is missing is not captured by CLIP similarity.





Fig. 10. Examples of images in the AI-Pastiche dataset not aligning well with their prompts, for content or style.

This highlights a critical limitation: CLIP’s perception is atomistic. It excels at verifying the presence of individual semantic concepts, but struggles to understand their relational structure, count, or stylistic execution as specified in a holistic prompt.

## 8 Conclusion

Our investigation into CLIP’s capacity to interpret artworks—encompassing both human-created and AI-generated images—reveals a model of considerable scope, yet one that remains limited in its ability to capture the depth and nuance characteristic of human aesthetic and contextual understanding. CLIP demonstrates competence in associating images with broad semantic categories and producing plausible textual descriptions. However, it exhibits consistent difficulty in handling more subjective dimensions of artistic analysis, including style, cultural context, and technical accuracy.

Critically, CLIP does not appear to possess a reliable ability to detect genuine artifacts of the generative process, such as anatomical distortions in hands or faces, or other structural inconsistencies commonly present in AI-generated imagery. These types of defects, while often salient to human observers, are not consistently reflected in CLIP’s visual embeddings.

In addition, a pronounced domain gap is observed when attempting to generalize stylistic classification across distinct datasets. While CLIP performs adequately within individual collections, both its accuracy and consistency decline when stylistic labels are transferred across datasets that differ in image distributions, curatorial practices, or annotation schemes. This limitation suggests a lack of robustness in CLIP’s abstraction of stylistic features beyond the specific contexts in which they are learned—a challenge that must be addressed if such models are to support applications such as cross-collection retrieval, cultural analysis, or curatorial decision-making.

At a more foundational level, this analysis raises important considerations about the structure of multimodal representations. Although CLIP’s joint embedding space helps bridge vision and language, this connection remains asymmetrical: text embeddings are inherently structured and interpretable, whereas image embeddings are more opaque and spatially entangled. As a result, assessing visual understanding through textual similarity may yield misleading conclusions—particularly in artistic domains, where meaning is often conveyed through non-verbal, ambiguous, or context-dependent visual signals. Elements such as texture, composition, gesture, and affect frequently resist straightforward linguistic encoding, exposing key limitations in current evaluation methodologies for multimodal models.

Looking forward, advancing vision–language systems will require a shift beyond surface-level alignment toward a deeper and more holistic model of visual perception. Such systems must be capable of reasoning about

imagery not only in terms of objects and styles but also with regard to historical reference, artistic intent, and visual narrative. Achieving this may require new forms of supervision that extend beyond caption-based training to include expert commentary, art historical metadata, culturally informed annotations, and dialogic multimodal interactions.

As AI becomes increasingly integrated into the production, interpretation, and dissemination of visual culture, it is essential to consider not only what models perceive but how they perceive—and whose perspectives are implicitly encoded. CLIP offers a compelling foundation, but it does not yet approximate human-level perception in the arts. Rather, it should be regarded as a partial and biased interpretive lens—one that can nonetheless yield valuable insights into both machine vision and human ways of seeing

## Data Availability

The code associated with this article, together with the dataset metadata and precomputed CLIP embeddings, is available in the following GitHub repository: [github.com/asperti/ART\\_though\\_CLIPs\\_eyes](https://github.com/asperti/ART_though_CLIPs_eyes).

## References

- [1] Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdiah Soleymani Baghshah. 2025. CLIP under the microscope: A fine-grained analysis of multi-object representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '25)*. IEEE, 9308–9317. Retrieved from [https://openaccess.thecvf.com/content/CVPR2025/html/Abbasi\\_CLIP\\_Under\\_the\\_Microscope\\_A\\_Fine-Grained\\_Analysis\\_of\\_Multi-Object\\_Representation\\_CVPR\\_2025\\_paper.html](https://openaccess.thecvf.com/content/CVPR2025/html/Abbasi_CLIP_Under_the_Microscope_A_Fine-Grained_Analysis_of_Multi-Object_Representation_CVPR_2025_paper.html)
- [2] Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. 2025. Vision-language models do not understand negation. arXiv:2501.09425. Retrieved from <https://arxiv.org/abs/2501.09425>
- [3] Andrea Asperti, Leonardo Dessi, Maria Chiara Tonetti, and Nico Wu. 2025. Does CLIP perceive art as we do? In *Proceedings of the 22nd International Conference on Content-Based Multimedia Indexing (CBMI '25)*. IEEE.
- [4] Andrea Asperti, Franky George, Tiberio Marras, Razvan Ciprian Stricescu, and Fabio Zanotti. 2025. A critical assessment of modern generative models' ability to replicate artistic styles. *Big Data Cogn. Comput.* 9, 9 (2025), 1–26. DOI: <https://doi.org/10.3390/bdcc9090231>
- [5] Andrea Asperti, Leonardo Naldi, and Salvatore Fiorilla. 2025. An investigation of the domain gap in CLIP-based person re-identification. *Sensors* 25, 2 (2025), 1–12. DOI: <https://doi.org/10.3390/s25020363>
- [6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL technical report. arXiv:2502.13923. Retrieved from <https://arxiv.org/abs/2502.13923>
- [7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV '21)*. IEEE, 1708–1718. DOI: <https://doi.org/10.1109/ICCV48922.2021.00175>
- [8] Alberto Baldreti, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2024. Composed image retrieval using contrastive learning and task-oriented CLIP-based features. *ACM Trans. Multim. Comput. Commun. Appl.* 20, 3 (2024), Article 62, 1–24. DOI: <https://doi.org/10.1145/3617597>
- [9] Yi Bin, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. 2024. GalleryGPT: Analyzing paintings with large multimodal models. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. ACM, 7734–7743. DOI: <https://doi.org/10.1145/3664647.3681656>
- [10] Tibor Bleidt, Sedigheh Eslami, and Gerard de Melo. 2024. ArtQuest: Countering hidden language biases in ArtVQA. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV '24)*. IEEE, 7311–7320. DOI: <https://doi.org/10.1109/WACV57701.2024.00716>
- [11] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. ShareGPT4V: Improving large multi-modal models with better captions. In *Proceedings of the 18th European Conference on Computer Vision (ECCV '24), Part XVII*. Springer Nature, 370–387. DOI: [https://doi.org/10.1007/978-3-031-72643-9\\_22](https://doi.org/10.1007/978-3-031-72643-9_22)
- [12] Marcos V. Conde and Kerem Turgutlu. 2021. CLIP-Art: Contrastive pre-training for fine-grained art classification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops '21)*. IEEE, 3956–3960. Retrieved from [https://openaccess.thecvf.com/content/CVPR2021W/CVFAD/html/Conde\\_CLIP-Art\\_Contrastive\\_Pre-Training\\_for\\_Fine-Grained\\_Art\\_Classification\\_CVPRW\\_2021\\_paper.html](https://openaccess.thecvf.com/content/CVPR2021W/CVFAD/html/Conde_CLIP-Art_Contrastive_Pre-Training_for_Fine-Grained_Art_Classification_CVPRW_2021_paper.html)
- [13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023 (NeurIPS '23)*. NIPS Foundation. Retrieved from [http://papers.nips.cc/paper\\_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html)
- [14] Alexey Dosovitskiy and Thomas Brox. 2016. Inverting visual representations with convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*. IEEE Computer Society, 4829–4837. DOI: <https://doi.org/10.1109/CVPR.2016.522>

- [15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. arXiv:2403.03206. Retrieved from <https://arxiv.org/abs/2403.03206>
- [16] Tarian Fu, Javier Conde, Gonzalo Martínez, Pedro Reviriego, Elena Merino-Gómez, and Fernando Moral. 2025. Artificial intelligence and misinformation in art: Can vision language models judge the hand or the machine behind the canvas? arXiv:2508.01408. Retrieved from <https://arxiv.org/abs/2508.01408>
- [17] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. CLIP-Adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.* 132, 2 (2024), 581–595. DOI: <https://doi.org/10.1007/s11263-023-01891-x>
- [18] Noa Garcia and George Vogiatzis. 2018. How to read paintings: Semantic art understanding with multi-modal retrieval. In *Proceedings of the Computer Vision - ECCV 2018 Workshops, Part II*. Springer Nature, 676–691. DOI: [https://doi.org/10.1007/978-3-030-11012-3\\_52](https://doi.org/10.1007/978-3-030-11012-3_52)
- [19] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. 2020. A dataset and baselines for visual question answering on art. In *Proceedings of the Computer Vision - ECCV 2020 Workshops, Part II*. Springer Nature, 92–108. DOI: [https://doi.org/10.1007/978-3-030-66096-3\\_8](https://doi.org/10.1007/978-3-030-66096-3_8)
- [20] Zeyi Huang, Andy Zhou, Zijian Lin, Mu Cai, Haohan Wang, and Yong Jae Lee. 2023. A sentence speaks a thousand images: Domain generalization through distilling CLIP with language guidance. In *IEEE/CVF International Conference on Computer Vision (ICCV '23)*. IEEE, 11651–11661. DOI: <https://doi.org/10.1109/ICCV51070.2023.01073>
- [21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 4904–4916.
- [22] Hamid Kazemi, Atoosa Malemir Chegini, Jonas Geiping, Soheil Feizi, and Tom Goldstein. 2024. What do we learn from inverting CLIP models? arXiv:2403.02580. Retrieved from <https://arxiv.org/abs/2403.02580>
- [23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML '23)*. PMLR, 19730–19742. Retrieved from <https://proceedings.mlr.press/v202/li23q.html>
- [24] Shengze Li, Jianjian Cao, Peng Ye, Yuhua Ding, Chongjun Tu, and Tao Chen. 2025. ClipSAM: CLIP and SAM collaboration for zero-shot anomaly segmentation. *Neurocomputing* 618 (2025), 129122. DOI: <https://doi.org/10.1016/j.neucom.2024.129122>
- [25] Siyuan Li, Li Sun, and Qingli Li. 2023. CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels. In *37th AAAI Conference on Artificial Intelligence (AAAI '23)*, *32th Conference on Innovative Applications of Artificial Intelligence (IAAI '23)*, *13th Symposium on Educational Advances in Artificial Intelligence (EAAI '23)*. AAAI, 1405–1413. DOI: <https://doi.org/10.1609/aaai.v37i1.25225>
- [26] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. 2022. The ArtBench dataset: Benchmarking generative models with artworks. arXiv:2206.11404. Retrieved from <https://arxiv.org/abs/2206.11404>
- [27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023 (NeurIPS '23)*. NIPS Foundation, 1–25. Retrieved from [http://papers.nips.cc/paper\\_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html)
- [28] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. 2023. More control for free! image synthesis with semantic diffusion guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV '23)*. IEEE, 289–299. DOI: <https://doi.org/10.1109/WACV56688.2023.00037>
- [29] Zhihang Liu, Qiang Huang, and Yingying Zhu. 2024. CLIP-based cross-level semantic interaction and recombination network for composed image retrieval. In *27th European Conference on Artificial Intelligence (ECAI '24) - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS '24)*. IOS Press, 704–711. DOI: <https://doi.org/10.3233/FAIA240552>
- [30] Vasyl Lytvyn, Roman Peleshchak, Ihor Rishnyak, Bohdan Kopach, and Yuriy Gal. 2024. Detection of similarity between images based on contrastive language-image pre-training neural network. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems. Volume I: Machine Learning Workshop (CEUR Workshop Proceedings)*. CEUR-WS.org, 94–104. Retrieved from <https://ceur-ws.org/Vol-3664/paper8.pdf>
- [31] Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*. IEEE Computer Society, 5188–5196. DOI: <https://doi.org/10.1109/CVPR.2015.7299155>
- [32] Leland McInnes and John Healy. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426. Retrieved from <http://arxiv.org/abs/1802.03426>
- [33] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML '22)*. PMLR, 16784–16804. Retrieved from <https://proceedings.mlr.press/v162/nichol22a.html>
- [34] National Gallery of Art. 2024. National Gallery of Art Open Data Program. Retrieved January 29, 2024 from <https://www.nga.gov/open-access-images/open-data.html>
- [35] Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, and Gao Huang. 2022. Contrastive language-image pre-training with knowledge graphs. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing*

- Systems 2022 (NeurIPS '22)*. NIPS Foundation, 22895–22910. Retrieved from [http://papers.nips.cc/paper\\_files/paper/2022/hash/904aac1c930c196f1c71533d4d9dc31a-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/904aac1c930c196f1c71533d4d9dc31a-Abstract-Conference.html)
- [36] Fang Peng, Xiaoshan Yang, Linhui Xiao, Yaowei Wang, and Changsheng Xu. 2024. SgVA-CLIP: Semantic-guided visual adapting of vision-language models for few-shot image classification. *IEEE Trans. Multimed.* 26 (2024), 3469–3480. DOI: <https://doi.org/10.1109/TMM.2023.3311646>
  - [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML '21), Virtual Event (Proceedings of Machine Learning Research)*. PMLR, 8748–8763. Retrieved from <http://proceedings.mlr.press/v139/radford21a.html> Accessed: 2025-02-13
  - [38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125. Retrieved from <https://arxiv.org/abs/2204.06125>
  - [39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 10684–10695.
  - [40] Vincenzo De Rosa, Fabrizio Guillaro, Giovanni Poggi, Davide Cozzolino, and Luisa Verdoliva. 2024. Exploring the adversarial robustness of CLIP for AI-generated image detection. In *IEEE International Workshop on Information Forensics and Security (WIFS '24)*. IEEE, 1–6. DOI: <https://doi.org/10.1109/WIFS61860.2024.10810719>
  - [41] Karsten Roth, Zeynep Akata, Dima Damen, Ivana Balazevic, and Olivier J. Hénaff. 2024. Context-aware multimodal pretraining. arXiv:2411.15099. Retrieved from <https://arxiv.org/abs/2411.15099>
  - [42] Ravidu Sui, Rammuni Silva, Ahmad Lotfi, Isibor Kennedy Ihianle, Golnaz Shahtahmassebi, and Jordan J. Bird. 2024. ArtBrain: An explainable end-to-end toolkit for classification and attribution of ai-generated art and style. arXiv:2412.01512. Retrieved from <https://arxiv.org/abs/2412.01512>
  - [43] Yogesh Surapaneni and Chakravarthy Bhagvati. 2024. Scene text image super-resolution with CLIP prior guidance. In *Proceedings of the 27th International Conference on Pattern Recognition (ICPR '24), Part XXXII*. Elsevier, 17–32. DOI: [https://doi.org/10.1007/978-3-031-78125-4\\_2](https://doi.org/10.1007/978-3-031-78125-4_2)
  - [44] Ashkan Taghipour, Morteza Ghahremani, Mohammed Bennamoun, Aref Miri Rekavandi, Zinuo Li, Hamid Laga, and Farid Boussaïd. 2024. Faster Image2Video generation: A closer look at CLIP image embedding's impact on spatio-temporal cross-attentions. arXiv:2407.19205. Retrieved from <https://arxiv.org/abs/2407.19205>
  - [45] Weijie Tu, Weijian Deng, and Tom Gedeon. 2024. Toward a holistic evaluation of robustness in CLIP models. arXiv:2410.01534. Retrieved from <https://arxiv.org/abs/2410.01534>
  - [46] Peng Wang, Dagang Li, Xuesi Hu, Yongmei Wang, and Youhua Zhang. 2025. CLIPMulti: Explore the performance of multimodal enhanced CLIP for zero-shot text classification. *Comput. Speech Lang.* 90 (2025), 101748. DOI: <https://doi.org/10.1016/j.csl.2024.101748>
  - [47] Yutaro Yamada, Yingting Tang, and Ilker Yildirim. 2022. When are lemons purple? The concept association bias of CLIP. arXiv:2212.12043. Retrieved from <https://arxiv.org/abs/2212.12043>
  - [48] Fan Yang, Nor Azman Ismail, Pang Yee Yong, and Alhuseen Omar Alsayed. 2025. CAMIR: Fine-tuning CLIP and multi-head cross-attention mechanism for multimodal image retrieval with sketch and text features. *Int. J. Multimed. Inf. Retr.* 14, 1 (2025), 2. DOI: <https://doi.org/10.1007/s13735-024-00352-6>
  - [49] Hairui Yang, Ning Wang, Haojie Li, Lei Wang, and Zhihui Wang. 2024. Application of CLIP for efficient zero-shot learning. *Multimed. Syst.* 30, 4 (2024), 219. DOI: <https://doi.org/10.1007/s00530-024-01414-9>
  - [50] Hongxu Yin, Pavlo Molchanov, José M. Álvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. 2020. Dreaming to distill: Data-free knowledge transfer via DeepInversion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '20)*. Computer Vision Foundation/IEEE, 8712–8721. DOI: <https://doi.org/10.1109/CVPR42600.2020.00874>
  - [51] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. 2023. Task residual for tuning vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '23)*. IEEE, Los Alamitos, CA, 10899–10909. DOI: <https://doi.org/10.1109/CVPR52729.2023.01049>
  - [52] Zeliang Zhang, Zhuo Liu, Mingqian Feng, and Chenliang Xu. 2024. Can CLIP count stars? An empirical study on quantity bias in CLIP. In *Findings of the Association for Computational Linguistics (EMNLP '24)*. Association for Computational Linguistics, 1081–1086. DOI: <https://doi.org/10.18653/V1/2024.FINDINGS-EMNLP.59>
  - [53] Kaiyang Zhou, Jingkan Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, 16795–16804. <https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01631>
  - [54] Kaiyang Zhou, Jingkan Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* 130 (2022), 2337–2348. Retrieved from <https://link.springer.com/article/10.1007/s11263-022-01653-1>
  - [55] Qiongyi Zhou, Changde Du, Shengpei Wang, and Huiguang He. 2024. CLIP-MUSED: CLIP-guided multi-subject visual neural information semantic decoding. In *The 12th International Conference on Learning Representations (ICLR '24)*. ICLR, 1–17. Retrieved from <https://openreview.net/forum?id=IKxL5zkssv>



- [56] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. 2023. Not all features matter: Enhancing few-shot CLIP with adaptive prior refinement. In *IEEE/CVF International Conference on Computer Vision (ICCV '23)*. IEEE Computer Society, Los Alamitos, CA, 2605–2615. DOI : <https://doi.org/10.1109/ICCV51070.2023.00246>
- [57] Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. 2024. Selective vision-language subspace projection for few-shot CLIP. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. ACM, 3848–3857. DOI : <https://doi.org/10.1145/3664647.3680885>

## Appendices

### A Example of an Artwork from the National Gallery

To illustrate the methodology adopted in our study, we selected a painting from the National Gallery of Art, Van Gogh's self-portrait. Since zero-shot CLIP accepts textual descriptions of no more than 77 tokens, we employed GPT-4o mini to generate a summary of the description of the painting, restricted to a maximum of 300 characters. The summarization process aimed to preserve essential information related to the subject and the stylistic characteristics of the artwork. This is the actual prompt used to generate the summaries:

Your goal is to summarize the following painting descriptions in 300 characters.

You will be provided the description of a painting, its subject, its style, and its period, and you will output a JSON object containing the following information:

{summary: string // at most 300 characters summary of the painting based on the painting description.}

The summary must retain information about the subject, style, and period.

In Figure A1, we present the painting; on the right, the official description provided by the National Gallery, and the summarized version produced using GPT-4o mini.



Fig. A1. Example of an artwork from the National Gallery.

#### National Gallery Description:

*Shown from the chest up, a man with short, orange hair and green-tinted, pale skin looks at us, wearing a vivid blue painter's smock in this vertical portrait painting. His smock and the background are painted with long, mostly parallel strokes of cobalt, azure, and lapis blue. His shoulders are angled to our left, and he looks at us from the corners of his blue eyes. He has a long, slightly bumped nose, and his lips are closed within a full, rust-orange beard. He holds a palette and paintbrushes in his left hand, in the lower left corner of the canvas. The background is painted with long brushstrokes that follow the contours of his head and torso to create an aura-like effect.*

#### Summarized version of the description:

*This Post-Impressionist portrait by Vincent van Gogh (1876–1900) depicts a man with orange hair and pale skin in a blue painter's smock. He gazes at the viewer with blue eyes, holding a palette and brushes. The background features long, parallel strokes of blue, enhancing the aura around him.*

The minimal summarization prompt described above was obtained after comparison with different and more elaborated versions, aimed at better exploiting CLIP capabilities according to common prescriptions [7, 21, 37]. In particular, we experimented with prompts encouraging embeddings that favor:

- concrete nouns (“man,” “table,” “sword”)
- observable relations (“holding,” “standing beside”)



- global visual properties (“dark background,” “warm colors”)
- Medium-level style cues (“oil painting,” “realistic,” “soft brushstrokes”)

The corresponding detailed prompt was the following:

You are a visual description compressor for paintings. Your task is to convert detailed painting descriptions into CLIP-compatible visual summaries for large-scale image–text matching.

Strict rules:

- Describe only what is visually observable in the painting.
- Do not infer symbolism, narrative meaning, emotion, or cultural context.
- Do not mention artist names, artwork titles, inscriptions, dates, periods, or provenance.
- Exclude fine facial anatomy unless visually dominant.
- Ignore metaphors and poetic language.
- Limit the output to approximately 60 tokens.

Prioritize (in order):

- Main subject(s) and count
- Pose, action, and interaction with objects
- Clothing, accessories, and materials
- Spatial layout and composition
- Dominant colors and lighting
- Background elements
- Generic painterly style cues (e.g. realistic, soft brushwork)

Compression rules:

- Merge redundant details.
- Use concrete nouns and adjectives.
- Prefer short visual phrases over full sentences.
- Avoid rare or specialized art-historical vocabulary.

Output format (mandatory):

- Single paragraph
- Comma-separated phrases
- No line breaks, no bullet points
- No introductory or concluding text

Your output must be stable, literal, and visually grounded.

This strategy produced summaries that were extremely dense in observable objects and attributes but structurally flat; an example is given in Table A1 (Summary 2). Although visually faithful, these outputs resembled perceptual inventories rather than natural captions. Empirically, the average image-to-summary cosine similarity remained around 0.5–0.6 across models, and Recall@K metrics showed a consistent degradation relative to the simpler summarization prompt. The result shows that CLIP is not optimized for matching images against pure perceptual inventories. Rather, it is trained contrastively on large-scale image–caption pairs collected from the Internet. These captions typically contain declared high-level visual categories (e.g., “portrait,” “religious painting,” “interior scene”), compositional abstractions, and stylistic descriptors expressed in natural language. When such abstractions are removed and replaced with low-level enumerations of objects and attributes, CLIP must implicitly reconstruct category structure from distributed cues.

To test whether the absence of high-level visual abstractions was responsible for the performance drop, we designed a third prompt that reintroduced such abstractions in a controlled manner, without allowing artist names, specific historical identities, or art-historical proper nouns. The precise prompt was:

Your goal is to summarize the following painting descriptions into CLIP-compatible visual summaries for image–text matching. You will be provided with:

- a subject title
- a style label
- a period label
- a detailed painting description

Rules:

- The summary must describe the painting as an image.
- Keep high-level visual categories inferred from subject, style, and period (e.g. portrait, religious panel painting, interior scene, Gothic style).
- Do NOT include artist names, dates, artwork titles, or specific historical identities.
- Prefer generic visual abstractions over proper nouns.
- The summary should read like a neutral museum caption, grounded in visible features.

Output format:

- One short declarative clause introducing the image type
- Followed by comma-separated visual phrases
- No line breaks, no bullet points
- No introductory or concluding commentary

The output must remain stable, literal, and visually grounded.

A typical output is given in Table [A1](#) (Summary 3).

Table A1. A Painting of the National Gallery of Art, Its Description and Three Different Summaries Generated through the Prompt Instructions Discussed in the Text



Young Lady.  
Sir Joshua Reynolds.

**Description:** A young girl with dark hair and pale skin, wearing a long white dress, kneels on the ground in a hilly, tree-lined landscape in this vertical portrait painting. The girl sits close to us, her body slightly angled to our left as she looks toward and reaches for a rose bush potted in a tall, brown urn along the left edge of the painting. She has pale gray eyes under faint brows and a delicate, pointed nose. Her round cheeks are flushed, and her pink bow lips are closed. Her chestnut-brown hair curls at the neck of her clothing, and bangs cover her forehead under a lacy white bonnet. Light glistens off her lace-edged black cape, which is tied at her throat. The full skirt of her white dress pillows like a cloud on the ground around her legs, and she has a wide, teal-blue sash around her waist. With her right arm, farther from us, she reaches out to touch a pink blossom with a gloved hand. The fingers of the butter-yellow gloves have been folded back. [...] Painted in smoky gray, ivory white, and pale, blush pink, the thin clouds against the topaz-blue sky fills the upper two-thirds of the picture. The painting is inscribed with gold paint in the lower right corner: “Lady Caroline Howard” and “Lady Cawdor.”

#### Summary 1

In this vertical portrait from 1776–1800 by Sir Joshua Reynolds, young Lady Caroline Howard kneels in a hilly landscape, dressed in a long white gown with a teal-blue sash. She reaches for a rose bush, her delicate features framed by a lacy bonnet. The scene is softly painted in pastel hues, capturing her innocence.

#### Summary 2

Young girl, dark hair, pale skin, kneeling, long white dress, reaching for rose bush in brown urn, slightly angled left, gray eyes, delicate nose, flushed cheeks, closed pink lips, chestnut curls, lace bonnet, black lace cape, teal-blue sash, gloved hand touching pink blossom, loose fist near knee, rolling hills, green trees, white sky, smoky gray.

#### Summary 3

This vertical portrait depicts a young girl kneeling in a hilly landscape, wearing a long white dress with a wide teal-blue sash, her dark hair adorned with a lacy white bonnet, reaching for a pink blossom on a rose bush in a brown urn, her pale skin and flushed cheeks contrasting with the soft colors of the surrounding scene.

This modification improved performance relative to the strictly perceptual summaries, confirming that explicit category-level structure facilitates alignment. However, the overall image-to-text similarity remained inferior to that obtained with the original summarization prompt.

This suggests that contemporary CLIP models are not well characterized as simple aggregators of local visual tokens. Instead, they appear better aligned with well-formed, structurally coherent captions that declare high-level visual categories early and explicitly. In other words, alignment quality depends not only on perceptual fidelity but also on the degree to which the text matches the statistical structure of caption data seen during training.

This observation has methodological implications: evaluating vision–language models using purely perceptual inventories may underestimate their retrieval capacity, while neutral, structured summaries may more faithfully reflect their learned alignment space.

## B Additional Examples of Mistakes in the Retrieval of Generative Prompts in the Case of AI-WikiArt

In this section, we investigate a distinct source of error in the AI-WikiArt image–prompt alignment task. Rather than focusing on semantically similar competing prompts, we concentrate on misclassified examples that exhibit the largest discrepancy between the cosine similarity assigned to the retrieved (incorrect) prompt and that of the ground-truth (correct) prompt. These cases are particularly informative because the gap in similarity scores suggests a structural breakdown in the alignment pipeline, rather than a marginal ranking error.

For AI-generated images, such severe misalignments typically arise from two principal sources. The first concerns the vision–language model itself. A model such as CLIP may fail to associate an image with its intended textual description due to limitations in perceptual sensitivity, stylistic discrimination, or higher-level semantic abstraction. In this scenario, the image may faithfully reflect the prompt, yet the embedding space does not adequately capture the relevant stylistic or contextual cues.

The second source of failure originates upstream, in the image generator. Here, the generator may not successfully translate the prompt into a visually coherent or stylistically consistent image. The resulting output can appear generic, internally inconsistent, or misaligned with the specified artistic intent. When this occurs, the vision–language model may in fact behave consistently: it assigns a low similarity score to the ground-truth prompt because the generated image no longer visually corresponds to it.

Disentangling these two failure modes is inherently challenging, as both ultimately manifest as reduced alignment scores. However, examining examples with large similarity gaps provides a diagnostic opportunity. In such cases, the magnitude of the discrepancy often reveals whether the breakdown is primarily perceptual (a limitation of the vision–language model) or generative (a failure of the image synthesis process).

In the examples shown in Figure B1, the issue does not stem from semantic proximity between competing prompts—unlike the cases discussed in Figure 6. Instead, the error can be traced to the generation stage: the model failed to produce an image that visually represents the true prompt. CLIP responds accordingly, assigning a very low similarity score to the ground-truth description. The system then selects an alternative “Predicted Prompt” that, while incorrect in terms of original intent, more accurately describes the actually generated image and therefore receives a higher similarity score. In this sense, the misclassification reflects a breakdown in generation rather than in retrieval.

Establishing a principled method to distinguish between these sources of error remains an open problem. Qualitative inspection can provide preliminary insights, but a reliable diagnostic framework would likely require systematic human evaluation or the development of advanced evaluative models capable of functioning as informed art critics. Such agents would need to reason about stylistic intent, visual coherence, aesthetic plausibility, and historical context—dimensions that extend beyond surface-level semantic alignment. Designing and probing these evaluative systems, and understanding whether they can meaningfully assess artistic fidelity and generative intent, constitutes a central direction for our future research.





**True Prompt** (Similarity = 0.111):

The image depicts the resurrection of Christ surrounded by praying angels and sleeping soldiers, created using Renaissance oil painting techniques.

**Predicted Prompt** (Similarity = 0.277):

The scene depicts a solemn group of people in historical attire gathered inside a church, with a young man standing prominently near a column, created using detailed oil painting techniques.

Generator: F-Lite

Style: High\_renaissance



**True Prompt** (Similarity = 0.136):

The image depicts two snow-covered haystacks in a winter landscape, created using Impressionist oil painting techniques with visible brushstrokes and a soft color palette.

**Predicted Prompt** (Similarity = 0.287):

The painting depicts a contemplative male artist holding a palette and brush in an outdoor setting, created using oil on canvas with expressive brushstrokes.

Generator: F-Lite

Style: Impressionism



**True Prompt** (Similarity = 0.146):

The image depicts a biblical scene of Saint Anthony being tormented by demons while surrounded by praying saints and onlookers, created using the fresco painting technique.

**Predicted Prompt** (Similarity = 0.295):

The scene depicts Jesus healing a blind man surrounded by onlookers in a classical architectural setting, created using Renaissance oil painting techniques.

Generator: FLUX

Style: Early\_renaissance

Fig. B1. Examples of misclassifications from the AI-WikiArt dataset where the image generator failed to adhere to the *True Prompt*. In all cases, CLIP correctly assigns a very low similarity score to the true prompt and a much higher score to a *Predicted Prompt* that more accurately describes the generated image's actual content.

### C Comparison of CLIP Architectures on the NGAD Art Style Classification Task for Different OpenAI Clip Versions

Table C1 reports the performance of different CLIP architectures on the NGAD style classification task. Among all evaluated models, ViT-L/14@336px achieved the highest scores, with a recall@1 of 0.457 and a recall@5 of 0.831. Overall, the results remain moderate across all models, reflecting the increased difficulty of style recognition compared to alignment with textual descriptions.

Table C1. Recall@k Scores for Art Style Recognition on NGAD

| Model          | rec@1        | rec@2        | rec@3        | rec@5        |
|----------------|--------------|--------------|--------------|--------------|
| RN50           | 0.354        | 0.546        | 0.662        | 0.801        |
| RN101          | 0.379        | 0.577        | 0.674        | 0.78         |
| RN50x4         | 0.344        | 0.504        | 0.611        | 0.772        |
| RN50x16        | 0.373        | 0.581        | 0.679        | 0.786        |
| RN50x64        | 0.343        | 0.516        | 0.627        | 0.766        |
| ViT-B/32       | 0.316        | 0.467        | 0.585        | 0.737        |
| ViT-B/16       | 0.349        | 0.506        | 0.622        | 0.766        |
| ViT-L/14       | 0.4          | 0.577        | 0.697        | 0.795        |
| ViT-L/14@336px | <b>0.457</b> | <b>0.632</b> | <b>0.716</b> | <b>0.831</b> |

Best results in bold.

### D CLIP Image Embedding Visualization

Figure D1 presents a 3D UMAP projection of the image embeddings from the NGAD, extracted using the CLIP ViT-L/14 model. Each point corresponds to a single painting, color-coded by its annotated artistic style.

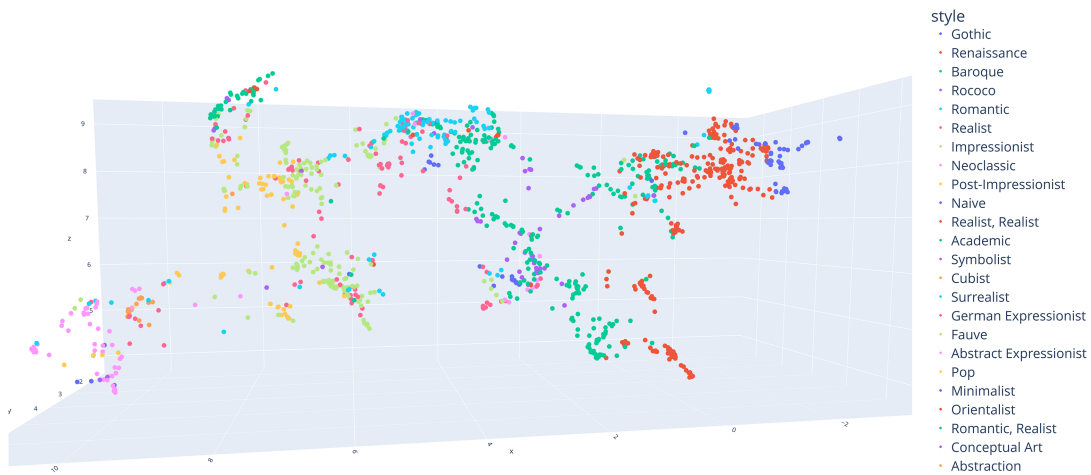


Fig. D1. 3D UMAP projection of image embeddings of NGAD extracted from CLIP.

The visualization reveals a clear semantic structure in the latent space: paintings with similar stylistic labels tend to form distinct clusters, suggesting that CLIP's visual encoder captures meaningful stylistic signals—even without explicit supervision for style.

This result is particularly noteworthy because the UMAP projection is unsupervised and agnostic to style labels. It is computed purely from the dominant eigenvectors of the image embedding space, meaning that no stylistic information was explicitly provided during dimensionality reduction. The emergence of style-based clusters therefore suggests that stylistic variance is one of the major synthetic components of the dataset and that CLIP encodes such variance in a way that aligns, at least partially, with human notions of artistic style.

Received 1 November 2025; revised 7 April 2026; accepted 18 April 2026