Do CLIP image embeddings reflect artistic style and chronology?

Rory Ashton

September 2025

Abstract

CLIP (Contrastive Language–Image Pre-training) is a large vision–language model trained on image–text pairs collected from publicly available web pages. Although it was not designed specifically for art, its embedding space is increasingly used to analyse artworks. This report investigates how much information about artistic style and the temporal development of an individual artist's work can be obtained from off-the-shelf CLIP image embeddings, without any additional training. Two small datasets are constructed: a Picasso-only corpus of 1081 paintings with year annotations (1890–1973), and a multi-artist dataset of 140 paintings spanning four twentieth-century styles (Abstract Expressionism, Analytical Cubism, Impressionism, and Surrealism). All images are embedded using a pretrained CLIP image encoder, and the resulting vectors are analysed using UMAP visualisations, k-nearest-neighbour (k-NN) classification, silhouette scores, and a neighbour-based chronology metric.

On the style dataset, a 5-NN classifier in CLIP space predicts the four style labels with 96% accuracy, and Analytical Cubism forms a particularly cohesive cluster, whereas overall silhouette scores are modest and Surrealism is only weakly separated. On the Picasso dataset, CLIP-nearest neighbours are typically 40–50% closer in year than random neighbours, indicating that the embedding reflects chronology to a meaningful but limited extent. Thumbnail-based UMAP views further show that local neighbourhoods are often dominated by shared subject matter, palette, and medium, with stylistic and temporal patterns emerging only where these visual properties align with style or period. Across both datasets, the findings suggest that CLIP arranges paintings mostly by visual and semantic content, while patterns related to style and temporal development are present but play a smaller role in determining their positions in CLIP's image space.

1. Introduction

CLIP (Contrastive Language–Image Pre-training) is a vision–language model from OpenAI that learns to align images with their associated natural-language descriptions (Radford et al., 2021). A key factor in its success is the massive scale of its training data, consisting of images with associated text harvested from publicly available web pages. These image–text pairs are processed independently by an image encoder and a text encoder to produce vector representations (embeddings) that live in a shared space. During training, the model is optimised with a contrastive loss so that each image is similar to its corresponding text and dissimilar to other texts in the batch, and likewise for each text with respect to the images.

Although CLIP is trained with supervision in the form of paired images and texts, it does not rely on a fixed set of class labels. Instead, each image is supervised by its accompanying free-form text description. This allows CLIP to generalise to new tasks by expressing them as text prompts, and to be sensitive to a wide range of semantic and stylistic attributes beyond those explicitly defined in standard closed-set classification models. This flexibility makes CLIP particularly appealing for the study of art, where nuance, ambiguity, and novelty are central.

Within this space, Asperti et al. (2025) investigated whether CLIP perceives artworks similarly to humans. They found that, while CLIP demonstrates impressive performance on many tasks, it tends to prioritise semantic content (what the painting depicts) and often overlooks stylistic attributes. They argue that this limitation is primarily a consequence of biases in CLIP's training data rather than an inherent weakness of the architecture itself. Ghildyal et al. (2025) used CLIP to predict scores for Wölfflin's five principles of art history, which cover aspects such as how depth is handled within an artwork and how clearly a subject is separated from its background. Consistent with Asperti et al., they also found that, out of the box, CLIP struggles to capture nuanced stylistic elements. However, by fine-tuning CLIP's image and text encoders on a supervised dataset annotated with Wölfflin scores, they achieved accurate predictions of these stylistic judgements, at the cost of reduced sensitivity to semantic content.

Against this background, this report asks two related questions about CLIP's off-the-shelf image embeddings. First, to what extent do they separate paintings by art-historical style, beyond the semantic content of what is depicted? Second, to what extent do they reflect the temporal development of a single artist's work, using the chronology of Picasso's paintings as a proxy?

To address these questions, this study constructs two small datasets: a Picasso-only corpus with year annotations, and a multi-artist dataset spanning four twentieth-century styles (Abstract Expressionism, Analytical Cubism, Impressionism, and Surrealism). All images are embedded using a pretrained CLIP image encoder, and the resulting vectors are analysed using two-dimensional UMAP visualisations, k-nearest-neighbour (k-NN) style classification, silhouette scores, and a neighbour-based chronology metric. On the style dataset, style labels can be predicted reliably with a simple 5-NN classifier, while silhouette scores reveal only modest cluster separation and especially weak separation for Surrealism. On the Picasso dataset, CLIP-nearest neighbours are typically substantially closer in year than random neighbours, indicating a meaningful but limited sensitivity to chronology. Together,

these analyses evaluate how effectively CLIP's image-space embeddings capture information about style and temporal development.

2. Dataset

The images used in this project were sourced from publicly available web pages. Picasso's paintings were downloaded from WikiArt (wikiart.org). The Picasso subset contains 1081 paintings with year annotations, spanning the period from 1890 to 1973. All works for which a year could be reliably obtained from WikiArt were included, so the subset reflects the range of Picasso's career as represented on the site.

For the style-comparison subset, a list of five artists was compiled for each style. The selection process began with informal research to identify artists widely recognised as central figures for each movement, followed by a check that a sufficient number of paintings were available online at acceptable resolution. No quantitative metric or fixed procedure was used to rank or select artists. Rather, the final list reflects a combination of historical prominence and practical availability of images. For each style, ten images were downloaded from the two artists with the most available works, and five images from each of the remaining three artists, yielding 35 images per style and 140 images in total. The styles considered were Abstract Expressionism, Analytical Cubism, Impressionism, and Surrealism. Table 1 lists the artists included for each style.

Abstract Expressionism	Franz Kline, Willem de Kooning, Lee Krasner, Jackson Pollock, Mark Rothko
Analytical Cubism	Georges Braque, Albert Gleizes, Juan Gris, Jean Metzinger, Pablo Picasso
Impressionism	Edgar Degas, Claude Monet, Camille Pissarro, Pierre-Auguste Renoir, Alfred Sisley
Surrealism	Salvador Dalí, Max Ernst, Frida Kahlo, René Magritte, Joan Miró

Table 1: Artists per style in the style-comparison subset

Because different movements are represented by different sets of artists with their characteristic subjects and compositions, the style subsets are not content-matched. In practice, this means that stylistic differences cannot be cleanly separated from differences in subject matter and composition, an issue returned to in the Limitations section.

Two CSV files were created to store metadata for the images. The Picasso metadata file contains the columns *filename*, *title*, and *year* (1081 rows). The style-comparison metadata file has the columns filename, title, year, and style (140 rows, 35 per style). The artist for each painting is inferred from the directory structure (style/artist/filename) when loading the dataset.

3. Methods

This section describes the processing pipeline used in the experiments. First, all images were embedded using a pretrained CLIP image encoder, producing a fixed-length vector representation for each artwork. These embeddings were then analysed in two complementary ways. For both the Picasso and style-comparison datasets, qualitative two-dimensional views of the embedding space were obtained using UMAP projections. On the style-comparison subset, the degree to which styles separate in CLIP space was quantified using k-NN classification accuracy and silhouette scores. On the Picasso subset, the extent to which the embedding reflects chronology was quantified using a neighbour-based year-gap metric that compares CLIP-nearest neighbours with randomly chosen neighbours.

3.1. Embedding strategy

All experiments use image embeddings extracted from a pretrained CLIP model from OpenAI, specifically the openai/clip-vit-base-patch32 variant. Only the image encoder was used; the text encoder was not loaded or applied at any stage.

Each artwork image was opened in RGB format and passed through CLIP's built-in preprocessing steps, using the CLIPProcessor from Hugging Face. This included resizing, centre-cropping, and normalising the image to match the format expected by the model. The processed image was then passed through the pretrained CLIP image encoder (with gradients turned off), which produced a 512-dimensional feature vector for that artwork, and this vector was then L2-normalised before further analysis.

One embedding was created per image and reused throughout the project. All nearest-neighbour comparisons use cosine distance between these L2-normalised embeddings, following standard CLIP practice. UMAP projections and silhouette scores were computed with Euclidean distance, using the default settings of the respective libraries, as described below. All experiments were implemented in Python using PyTorch, the Hugging Face transformers library, umap-learn, scikit-learn, Matplotlib, and Plotly.

3.2. UMAP visualisation

To obtain a qualitative view of how CLIP organises the artworks, the 512-dimensional image embeddings were projected into two dimensions using Uniform Manifold Approximation and Projection (UMAP). For each experiment (Picasso-only and style-comparison), UMAP was fitted on the full matrix of CLIP embeddings returned by the image encoder. The umap-learn implementation was used with n_neighbors = 15 and min_dist = 0.1 (its default parameters) and a fixed random_state = 42, which corresponds to Euclidean distance in the original embedding space. Label information (year or style) was excluded from the UMAP fitting and used only afterwards to colour points in the plots. UMAP projections are used only for qualitative visualisation, and all quantitative metrics are computed in the original 512-dimensional embedding space.

For the Picasso subset, each point in the UMAP space corresponds to a single painting. A two-dimensional UMAP projection of the CLIP embeddings was first computed and then

visualised in two ways. In the static Matplotlib scatter plot, each point is coloured according to the painting's year of creation, with years first rescaled using min–max normalisation so that the full range maps onto the colour bar. In addition, an interactive Plotly scatter plot was generated in which each point is replaced by a small thumbnail of the corresponding painting, allowing direct visual inspection of local neighbourhoods in the embedding. For the style-comparison subset, UMAP was again fitted on the CLIP embeddings and the result was visualised as a scatter plot in which each point is coloured by its style label (Abstract Expressionism, Analytical Cubism, Impressionism, or Surrealism).

3.3. k-NN & silhouette metrics (style dataset)

To quantify how well CLIP's embedding space separates artworks by style, two simple cluster-quality measures were computed on the style-comparison dataset using k-NN classification accuracy and silhouette scores.

For k-NN, each artwork was represented by its CLIP image embedding, and the style labels (Abstract Expressionism, Analytical Cubism, Impressionism, Surrealism) were used as class labels. Using scikit-learn's NearestNeighbors with cosine distance, a 5-nearest-neighbour index was fitted on all embeddings. For each image, its five nearest neighbours (excluding the image itself) were retrieved and it was assigned the majority style among those neighbours, with a fixed tie-breaking rule. The k-NN accuracy is defined as the proportion of images whose predicted style matches their true style label. Per-style accuracies are also reported by averaging this correctness indicator within each style.

In addition to k-NN accuracy, silhouette scores were used to measure how clearly the styles are separated in the CLIP embedding space. The known style labels are treated as cluster memberships. For each image, scikit-learn's silhouette_samples function computes a score between –1 and 1 that compares its average distance to images of the same style with its average distance to images of other styles (using Euclidean distance in the CLIP embedding space). Scores close to 1 indicate that the image is much closer to its own style cluster than to other styles, scores near 0 indicate overlapping clusters, and negative scores indicate that the image is, on average, closer to a different style than to its own. Both the overall mean silhouette score across all images and the mean silhouette score per style are reported. In this setting, k-NN accuracy and silhouette scores are used simply to describe how clustered the data are in the full dataset, rather than to measure how well a model generalises to new, unseen data.

3.4. Chronology metric (Picasso dataset)

For the Picasso subset, chronology was quantified using a neighbour-based metric. First, cosine distances were computed between every pair of Picasso embeddings. For each painting, its k-nearest neighbours in the CLIP space (excluding the painting itself) were identified, and the average absolute difference in year between the painting and its neighbours was calculated. Averaging this value over all paintings gives the typical year gap between CLIP-nearest neighbours. As a baseline, the same procedure was repeated but with k randomly chosen other paintings in place of nearest neighbours, yielding a typical year gap for random neighbours. For each reference work, a set of random neighbours was selected from the other paintings, without selecting the same painting twice within that neighbourhood. For several values of k (3, 5, 10, 20), the mean year gap to nearest

neighbours and to random neighbours is reported, together with the percentage reduction in year gap, with these values chosen to probe both very local neighbourhoods and more global structure in the embedding. Higher values of this percentage indicate that CLIP tends to place temporally closer works nearer to each other in the embedding space than would be expected by chance.

4. Results

4.1. Style comparison UMAP

Figure 1 shows the two-dimensional UMAP projection of the CLIP embeddings for the style-comparison dataset, with points coloured by style label. Although the UMAP fit did not use the style labels, the four styles form clearly separated, compact clusters. Abstract Expressionism and Impressionism occupy two well-isolated regions on the left and upper part of the plot, while Analytical Cubism and Surrealism form two distinct but neighbouring clusters on the right.

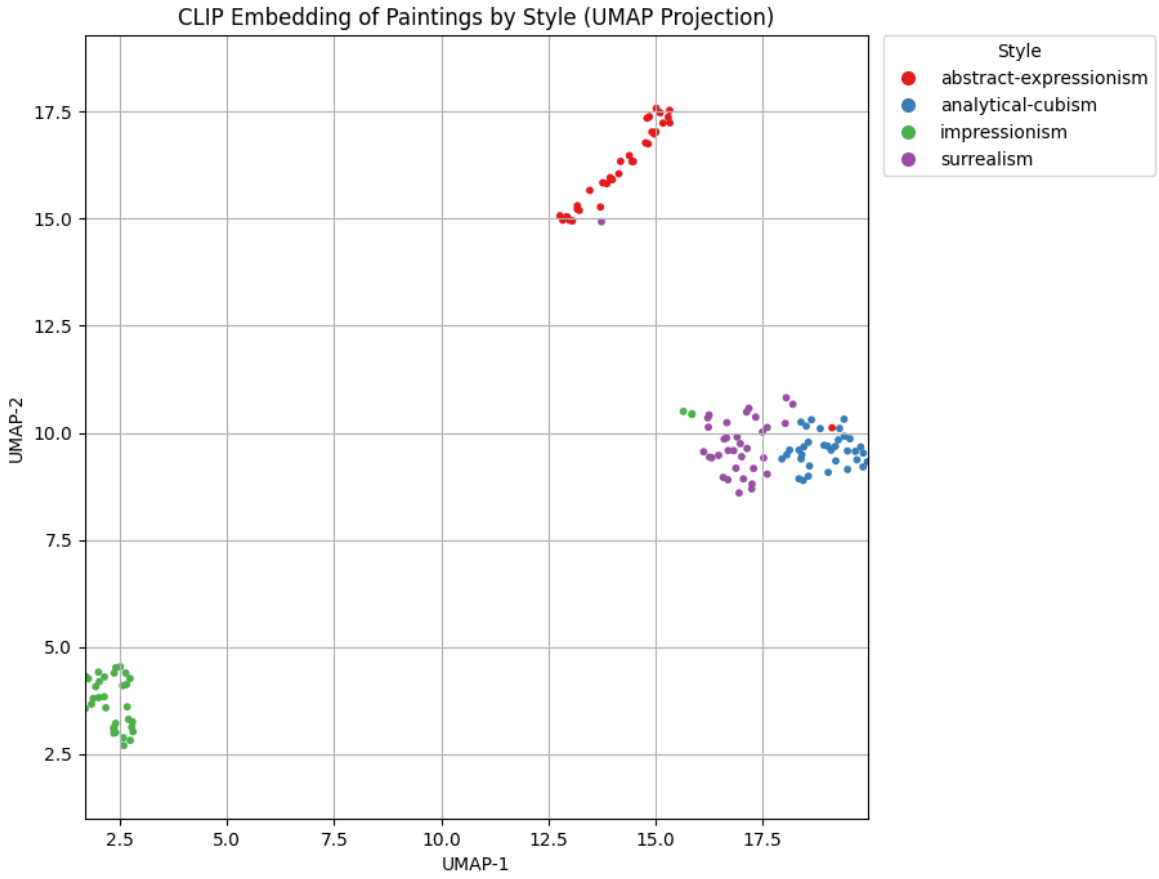

Figure 1

In this projection, the four styles appear as compact clusters with very little visible overlap at this level of visualisation. Analytical Cubism and Surrealism are located close to each other on the right-hand side of the manifold, whereas Impressionism and Abstract Expressionism occupy more distant regions. Section 4.4 quantifies how well these styles are separated in the original CLIP embedding space.

4.2. Picasso chronology UMAP

Figures 2 and 3 show the UMAP projection of the CLIP embeddings for Picasso's paintings. In both plots, the projection is fitted without year labels. The colour of each point indicates the (normalised) year of creation, and in Figure 3 each point is replaced by a thumbnail of the corresponding painting.

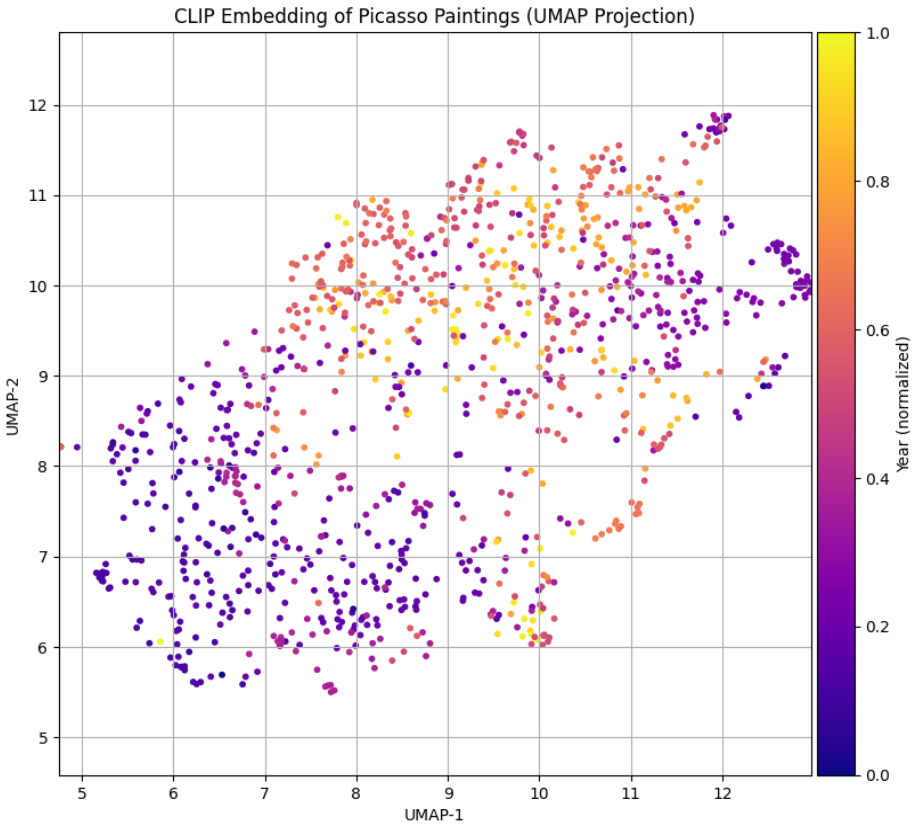

Figure 2

The point cloud in Figure 2 forms a roughly elongated structure, and the colour gradient reveals a large-scale temporal pattern. Earlier works (darker colours) are concentrated towards one end of the manifold, while later works (lighter colours) appear more frequently towards the opposite end. Moving along the main diagonal of the plot, the colours tend to vary gradually rather than randomly, indicating that nearby points in the CLIP embedding space are often close in time. At the same time, the colouring is far from perfectly ordered, with pockets of early and late works mixed together in many areas and paintings from similar years appearing in quite different regions.

Figure 3 shows the corresponding thumbnail-based UMAP visualisation. Here each point is replaced by a small image of the painting, with the border colour again indicating the normalised year. When zooming into local neighbourhoods in Figure 3, groups of paintings that are close in the UMAP space tend to share visible properties such as colour palette, composition, subject matter, or medium.

In one zoomed region (Figure 3a), the neighbourhood consists almost entirely of portraits, and neighbouring paintings move from more naturalistic, conventionally modelled faces to later, more stylised and fragmented ones, with the year colours also changing gradually across this local area. In another zoomed region (Figure 3b), the neighbourhood is dominated by monochrome drawings and sketch-like works, many of them animal studies, which span a range of years.

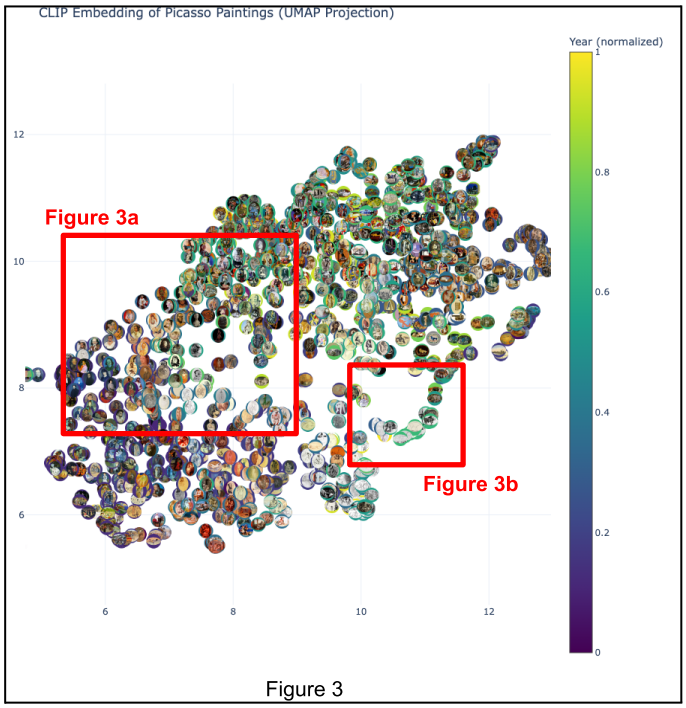

Figure 3

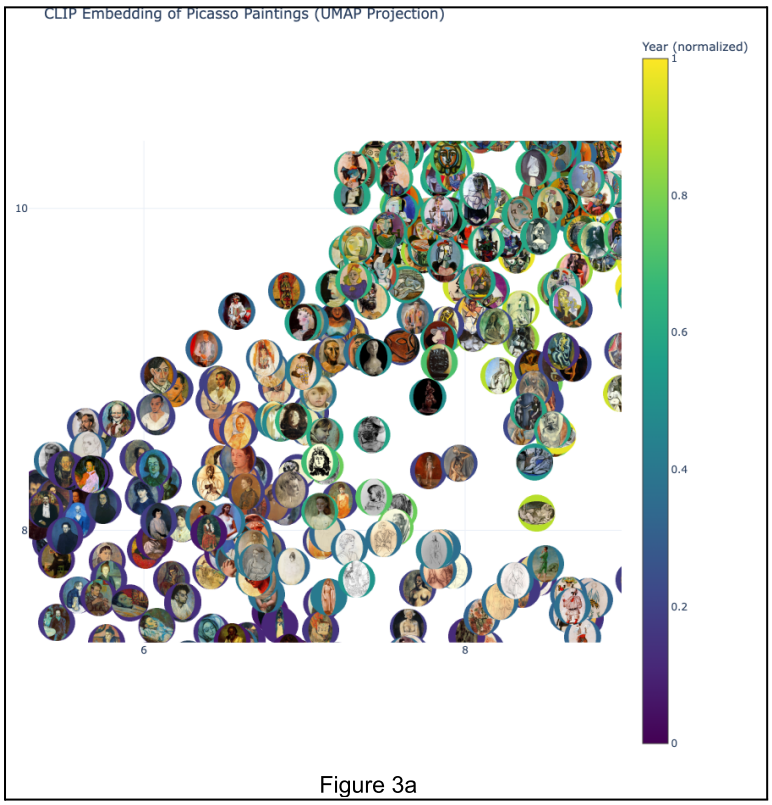

Figure 3a

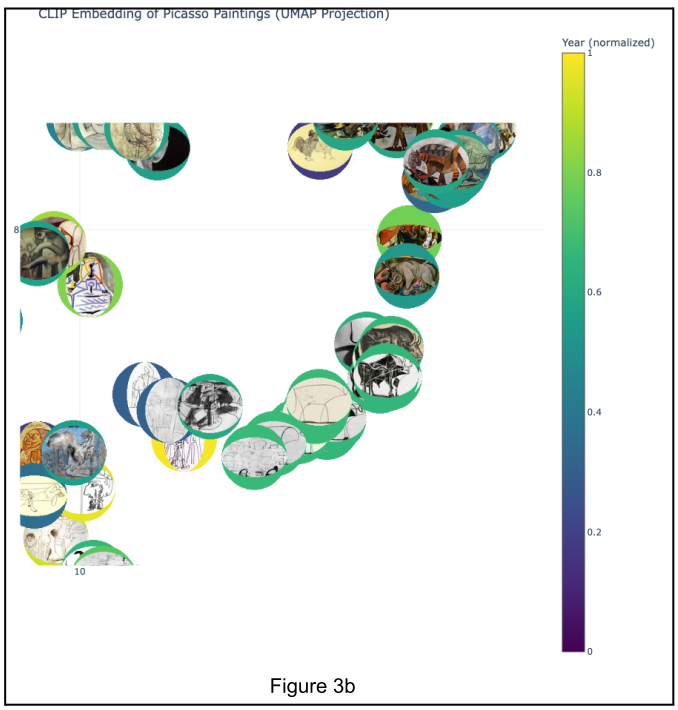

Figure 3b

4.3. Picasso chronology metric

To quantify how strongly CLIP's embedding space reflects the chronology of Picasso's work, the analysis compares the average year differences between each painting and its neighbours in CLIP space with the corresponding gaps for random neighbours. Table 2 shows that, for very local neighbourhoods (k = 3), CLIP neighbours are on average about 10 years apart, compared with around 23 years for random neighbours (a reduction of roughly 57%). As k increases to 5, 10 and 20, the neighbour gaps grow but remain around 40–53% smaller than the random baseline, indicating that nearby points in the CLIP embedding are typically much closer in time than would be expected by chance.

k	Mean Year Gap: CLIP Neighbours	Mean Year Gap: Random Neighbours	Year Gap Reduction vs Random (%)
3	10.012643	23.250385	56.935584
5	10.810731	23.032377	53.062897
10	12.324514	23.121462	46.696647
20	13.837234	23.333673	40.698430

Table 2

Figure 4 shows the distribution of per-painting average year gaps for k = 10. The curve for CLIP neighbours is shifted towards smaller gaps (mean **12.3** years vs **23.2** years for random), but the two distributions still overlap, and there is a clear subset of paintings whose neighbours are 20–30 years apart.

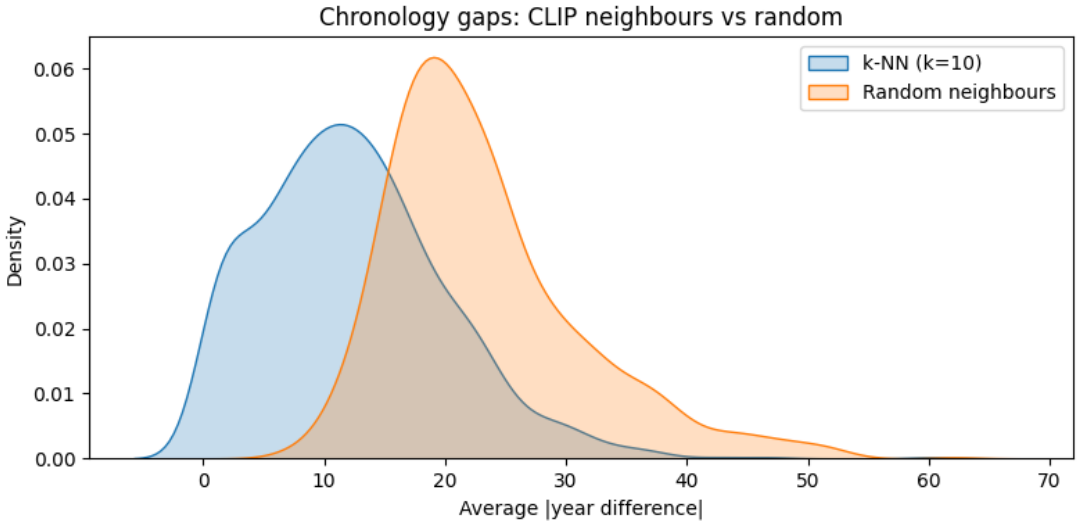

Figure 4

These results show that year gaps to CLIP neighbours are typically smaller than to random neighbours across all values of k considered.

4.4. Style separability with k-NN accuracy & silhouette metrics

Table 3 summarises the k-NN classification accuracy and silhouette scores for the style-comparison dataset. Overall, a 5-nearest-neighbour classifier in CLIP space correctly recovers the style label for about 96% of paintings. Analytical Cubism is classified perfectly

in this setting, while Abstract Expressionism and Impressionism are also very reliable. Surrealism has the lowest accuracy, with a noticeable but still moderate drop compared with the other three styles.

Style	k-NN accuracy (k=5)	Mean silhouette
All styles	0.957	0.137
Analytical Cubism	1.000	0.337
Abstract Expressionism	0.971	0.125
Impressionism	0.971	0.095
Surrealism	0.886	-0.009

Table 3

The silhouette scores in Table 3 give a more cautious view of cluster separation in the original CLIP embedding. The overall mean silhouette is modestly positive, indicating that the four styles are separated but not cleanly partitioned. Analytical Cubism again stands out with the highest silhouette value, meaning that those paintings tend to be much closer to others of the same style than to paintings of different styles. Abstract Expressionism and Impressionism have lower but still positive silhouettes, whereas Surrealism has a slightly negative mean silhouette, implying that many Surrealist works lie closer to other styles than to the centre of their own cluster.

5. Discussion

5.1. Style in CLIP's embedding

The style-comparison experiments suggest that CLIP's image embedding space contains substantial, but not uniform, information about art-historical styles. On the style-comparison dataset, the evidence is strong: a simple 5-nearest-neighbour classifier in the CLIP embedding space recovers the four style labels with around 96% accuracy overall, and Analytical Cubism in particular is almost perfectly separated (Table 3). In the UMAP projection (Figure 1), each style forms a compact region with little visible overlap, and Analytical Cubism appears especially cohesive. The main exception is Surrealism, which not only has the lowest k-NN accuracy but also a slightly negative mean silhouette, and occupies an intermediate position between the Analytical Cubist and Impressionist regions.

Overall, the k-NN and silhouette results suggest that some styles form tight, well-defined groups in CLIP's image space, while others gradually blend into nearby styles. Analytical Cubism stands out as the clearest group, with both high classification accuracy and a high silhouette score, which suggests that its recurring visual patterns match well with the features CLIP has learned from web-scale training data. By contrast, Surrealism is much more spread out in the embedding space: many Surrealist paintings are closer (in Euclidean distance) to works from other styles than to other Surrealist works. This may reflect the visual diversity of Surrealism itself, as well as overlaps in subject matter and composition with neighbouring movements.

These findings fit naturally alongside prior work. Asperti et al. (2025) and Ghildyal et al. (2025) report that off-the-shelf CLIP tends to prioritise semantic content (what is depicted) over stylistic nuance, and that fine-tuning on art-specific labels or Wölfflin scores is needed to obtain reliable style predictions. The present results are in line with this picture and suggest that broad style labels can often be recovered from CLIP's fixed embeddings, especially when a movement has a strong and distinctive visual appearance, but that the embedding does not separate all styles cleanly.

5.2. Chronology and Picasso

By contrast, the Picasso experiments focus on temporal development. In the UMAP projection for the Picasso subset (Figure 2), there is a clear early-to-late drift in the embedding space, with earlier works tending to cluster toward one side and later works toward the other, and colours changing gradually along the main diagonal rather than at random. This suggests a large-scale organisation in which paintings produced closer in time are often located nearer to each other in the low-dimensional projection, even though UMAP itself is fitted without access to year labels.

This pattern is measured more directly by the neighbour-based chronology metric in Table 2, which is computed in the original 512-dimensional space. Across a range of neighbourhood sizes, neighbours in CLIP space are between 40% and 50% closer in year than randomly chosen neighbours. Figure 4 compares the full distributions of year gaps. The CLIP curve is clearly shifted towards smaller gaps than the random curve, which means that paintings that lie near each other in CLIP space are, on average, painted closer together in time than would be expected by chance.

At the same time, the link between position in CLIP space and year is limited, and seems weaker than the influence of visual similarity. The chronology distributions in Figure 4 include many large year gaps, and there is a clear subset of paintings whose CLIP neighbours are 20–30 years apart, while the UMAP plots in Figures 2 and 3 show regions where early and late works appear side by side. In the thumbnail-based visualisation, local neighbourhoods such as Figure 3a and Figure 3b are dominated by shared properties of subject matter, colour palette, composition, or medium. Works that are all portraits or all monochrome animal studies end up adjacent even when they span decades, while paintings from similar years can be scattered into different regions if their surface appearance diverges.

One way to read these results is that CLIP arranges Picasso's paintings mainly according to the visual and semantic features it has learned from web data, such as subject matter, layout, and overall appearance, with time playing a smaller role on top of that. When particular periods in Picasso's career line up with distinctive and fairly stable visual styles, the temporal structure shows up more clearly in the embedding, which helps explain the reduced year gaps. When subject matter and surface appearance cut across different periods, the embedding behaves more like a content-based similarity space than a timeline of his career.

5.3. Limitations and implications

There are several limitations that constrain how far these conclusions can be pushed. First, the datasets are small and not balanced in a way that disentangles style from other factors. For each style, two artists provide the majority of examples, so it is difficult to know how much of the separability reflects the style label and how much reflects individual artist signatures.

Second, the style labels used here are defined at the level of broad movements rather than individual works. Abstract Expressionism, Analytical Cubism, Impressionism, and Surrealism are historically useful categories, but many individual works occupy ambiguous positions between movements or combine elements of several at once. In the experiments, each painting is forced into a single style label, and the evaluation assumes this assignment is unambiguous, although stylistic boundaries at the level of individual works can be much less clear-cut, even for paintings chosen as clear examples of their respective movements. The slightly negative silhouette for Surrealism may therefore reflect not only CLIP's limitations, but also the fact that some of the Surrealist works in this selection are visually close to neighbouring styles in this representation.

Third, the experiments do not control for semantic content. Different styles in the dataset have different typical subjects and compositions, and no attempt is made to match, for example, portraits across all four styles. As a result, some of the apparent style separability in Table 3 may reflect differences in subject matter and composition rather than stylistic factors alone. Likewise, in the Picasso chronology analysis, the metric cannot distinguish effects of changing subject matter from effects of time.

Finally, only a single CLIP variant (openai/clip-vit-base-patch32) and a single representation (the final image embedding) are considered. Other CLIP architectures, different layers, or models trained specifically on art datasets might show different behaviour. The present study therefore gives a snapshot of how one general-purpose vision–language model "sees" art, rather than a definitive characterisation of all large vision models.

These limitations mean that the findings should be interpreted as suggestive rather than definitive: they indicate patterns in how CLIP represents style and chronology for these particular datasets, but they do not fully separate stylistic, semantic, and artist-specific effects. Even so, the combination of high style recoverability for some movements and a measurable chronology signal for Picasso suggests that off-the-shelf CLIP embeddings already encode a non-trivial amount of art-historical structure, albeit in a way that is entangled with content and surface appearance.

5.4. Directions for future work

Several of the limitations above point directly to possible extensions of this work. A straightforward next step would be to construct content-controlled subsets, for example by selecting portraits, still lifes, or landscape scenes across multiple styles, and then repeating the k-NN and silhouette analyses. This would help isolate how much of the current separability is truly stylistic, and how much is due to differences in subject or composition.

For Picasso, it would be interesting to replace raw year labels with more art-historically meaningful period labels (e.g. Blue Period, Rose Period, early Cubism, late work) and test whether CLIP distinguishes those periods more clearly than it does calendar time. This could

be combined with the neighbour-based chronology metric to see whether CLIP's embedding is better aligned with standard art-historical periods than with a simple chronological ordering by year.

A second direction would be to compare CLIP with models trained directly on art, or to fine-tune CLIP on art-specific annotations (such as style labels or Wölfflin scores) and then re-evaluate the same metrics. This would connect the present, purely observational study to the fine-tuning results reported by Ghildyal et al., and could shed light on what is gained and lost when CLIP is adapted more aggressively to art.

Finally, the text side of CLIP has not been used here. Probing the text encoder with prompts such as "an abstract expressionist painting", "an analytical cubist portrait", or "a late Picasso drawing" and measuring retrieval quality on the same datasets would make it possible to compare how style is represented in CLIP's image and text spaces. Pursuing these directions would help develop this study into a more systematic account of how modern vision–language models represent style, period, and content in art.

6. Conclusion

This project used CLIP image embeddings to probe how a large vision–language model organises artworks by style and over time. On a small, curated dataset of twentieth-century movements, style labels are highly recoverable with a simple 5-NN classifier, and on a 1081-painting Picasso corpus, neighbours in CLIP space are typically 40–50% closer in year than random neighbours. These results indicate that large-scale stylistic and chronological structure can emerge in CLIP's image space even when the model is used off the shelf, without any additional training on art or art-historical labels.

Allowing for these limitations, the findings suggest a particular picture of how CLIP represents art. Rather than structuring its representation around style or period, CLIP seems to organise artworks primarily by visual similarity in subject matter, surface appearance, colour, and composition. Art-historical categories such as Analytical Cubism, Impressionism, or early versus late Picasso become more apparent when they line up with stable visual and semantic patterns and with systematic changes over time, but they are not treated as fundamental organising principles. For tasks that require nuanced stylistic or period-sensitive judgement, such as those based on Wölfflin's principles, some form of fine-tuning or explicit supervision therefore still seems necessary, and may come with trade-offs in semantic sensitivity as previous work has suggested. The results are in line with earlier work suggesting that CLIP is mainly driven by what is depicted, but they also show that its fixed image embeddings still contain a noticeable amount of art-historical structure.

The analyses here are necessarily narrow, focusing on a single CLIP variant and two modest datasets. Nonetheless, they illustrate how large vision–language models can be probed to uncover art-historical structure in their representations using simple nearest-neighbour and clustering analyses on fixed embeddings, and they indicate several directions for further work, including building content-controlled datasets, incorporating period labels, comparing different architectures and training regimes, and bringing the text encoder into the picture. Compared with studies that fine-tune CLIP on art-specific labels, this project treats CLIP purely as a feature extractor and asks how far style and chronology can be recovered from

its existing image space, a perspective that is directly relevant to using large pretrained models as general-purpose tools in computational art history.

References

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021) ‘Learning Transferable Visual Models From Natural Language Supervision’, arXiv preprint arXiv:2103.00020.

Asperti, A., Dessì, L., Tonetti, M. C. & Wu, N. (2025) ‘Does CLIP perceive art the same way we do?’, arXiv preprint arXiv:2505.05229.

Ghildyal, A., Wang, L.-Y. & Liu, F. (2025) ‘WP-CLIP: Leveraging CLIP to Predict Wölfflin’s Principles in Visual Art’, arXiv preprint arXiv:2508.12668.