Do CLIP style predictions extend to unseen artists?

1. Introduction

The main report asked whether off-the-shelf CLIP image embeddings separate paintings by art-historical style and reflect the temporal development of Picasso's work. Two small datasets were used in this study. One is a Picasso-only corpus with year labels, and the other is a style dataset of 140 paintings from four twentieth-century movements (Abstract Expressionism, Analytical Cubism, Impressionism, and Surrealism). All images were embedded with a pretrained CLIP image encoder, and the resulting vectors were analysed using UMAP, 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 correctly recovered the style label for around 96% of paintings, with perfect accuracy for Analytical Cubism and slightly lower accuracy for Surrealism. In that analysis, k-NN accuracy measured how tightly the dataset clusters by style when all paintings are allowed to act as neighbours. It did not directly address how well style predictions would transfer to genuinely new data.

This short report focuses on a more specific question: if style is predicted using CLIP embeddings and k-NN, does the approach still work for paintings by artists whose works are never used as neighbours? This is a stricter setting than the original leave-one-out evaluation, where the neighbour pool can include many works by the same artist. If accuracy remains high when same-artist neighbours are removed, this would indicate that CLIP captures style information that generalises across artists. A large drop in accuracy would instead suggest that the original result relied more on recognising particular artists and their characteristic subjects or compositions.

2. Method

2.1. Style dataset

The style dataset contains 140 paintings in total, with 35 images for each of four movements: Abstract Expressionism, Analytical Cubism, Impressionism, and Surrealism.

For each style, five artists were selected based on art-historical prominence and image availability. Two artists per style contributed ten paintings each, and three contributed five paintings each, giving 35 paintings per style overall. Each painting has a style label (one of the four movements), an artist name, and a 512-dimensional CLIP image embedding.

All images were embedded using a pretrained CLIP image encoder (no fine-tuning), then L2-normalised. One embedding was computed per image and reused across all analyses. Cosine distance was used for nearest-neighbour computations.

2.2. Original k-NN style accuracy

The original style analysis constructed a 5-NN index on all 140 embeddings using cosine distance. For each painting:

1. Its five nearest neighbours (excluding itself) were retrieved.
2. The predicted style was the majority style among these neighbours.
3. A fixed rule was used to break ties.

Per-style accuracy is the proportion of paintings in that style whose predicted label matches the true label. Overall accuracy is the proportion of correctly classified paintings across all four styles. In this original setting, every painting can use other works by the same artist as neighbours. This leads to high accuracies:

- Overall accuracy = 0.957
- Per-style accuracies:
    - Abstract Expressionism = 0.971
    - Analytical Cubism = 1.000
    - Impressionism = 0.971
    - Surrealism = 0.886

These numbers provide the baseline for the unseen-artist experiment.

2.3. Unseen-artist neighbour protocol (single run)

The unseen-artist experiment keeps CLIP frozen and reuses the same embeddings, but restricts which paintings are allowed to act as neighbours in the k-NN classifier. The aim is to simulate style prediction for artists who are not represented in the neighbour pool.

A single held-out split is constructed as follows. For each style, one artist is chosen at random to be held out. All paintings by these four artists (one per style) form the set of query paintings, giving 20 queries in total. The remaining 120 paintings, from the other sixteen artists, form the reference set used as neighbours.

A 5-NN index is fitted on the reference embeddings only. For each query painting, its five nearest neighbours are drawn from the reference set, and the painting is assigned the majority style among those neighbours. Overall and per-style accuracies are then computed on the 20 query paintings.

Apart from this restriction on the neighbour pool, all settings are identical to the original experiment: same CLIP model, embeddings, normalisation, distance metric, and value of k.

2.4. Multiple unseen-artist runs

Because the dataset is small, the result of a single random choice of held-out artists can be noisy. To obtain more stable estimates, the unseen-artist split is repeated 10 times with different random seeds. In each run, a new set of held-out artists (one per style) is sampled, the corresponding query and reference sets are formed, and 5-NN style accuracy is computed as in Section 2.3.

For each style and for the overall metric, the mean and standard deviation of accuracy across the 10 runs are then calculated. A summary table is constructed that includes the original 5-NN accuracies (overall and per style), the mean unseen-artist accuracies, and the corresponding standard deviations.

3. Results

3.1. Overall accuracy

Table 1 reports the overall k-NN style accuracy for the original evaluation and for the unseen-artist evaluation averaged over 10 runs.

Setting	Overall accuracy
Original (full neighbour pool)	0.957
Unseen artists (10-run mean ± s.d.)	0.922 ± 0.035

Table 1: Overall 5-NN style accuracy in CLIP space

In the original configuration, a 5-NN classifier in CLIP space correctly recovers the style label for about **95.7%** of paintings. When neighbours are restricted to paintings by other artists, the mean overall accuracy over held-out artists is **0.9218**, with a standard deviation of **0.0353**. The drop from 0.957 to about 0.922 is modest, and the classifier remains highly accurate under the stricter condition. This supports the view that CLIP's embeddings contain style information that generalises to new artists, at least within this small curated dataset.

3.2. Per-style effects

Table 2 shows per-style accuracies under the two evaluation protocols.

Setting	Abstract Expressionism	Analytical Cubism	Impressionism	Surrealism
Original	0.971	1.000	0.971	0.886
Unseen artists (mean)	0.960	1.000	1.000	0.700

Table 2: Per-style 5-NN accuracy (original vs unseen artists)

Figure 1 shows the same comparison visually, with separate bars for the original and unseen-artist settings for each style.

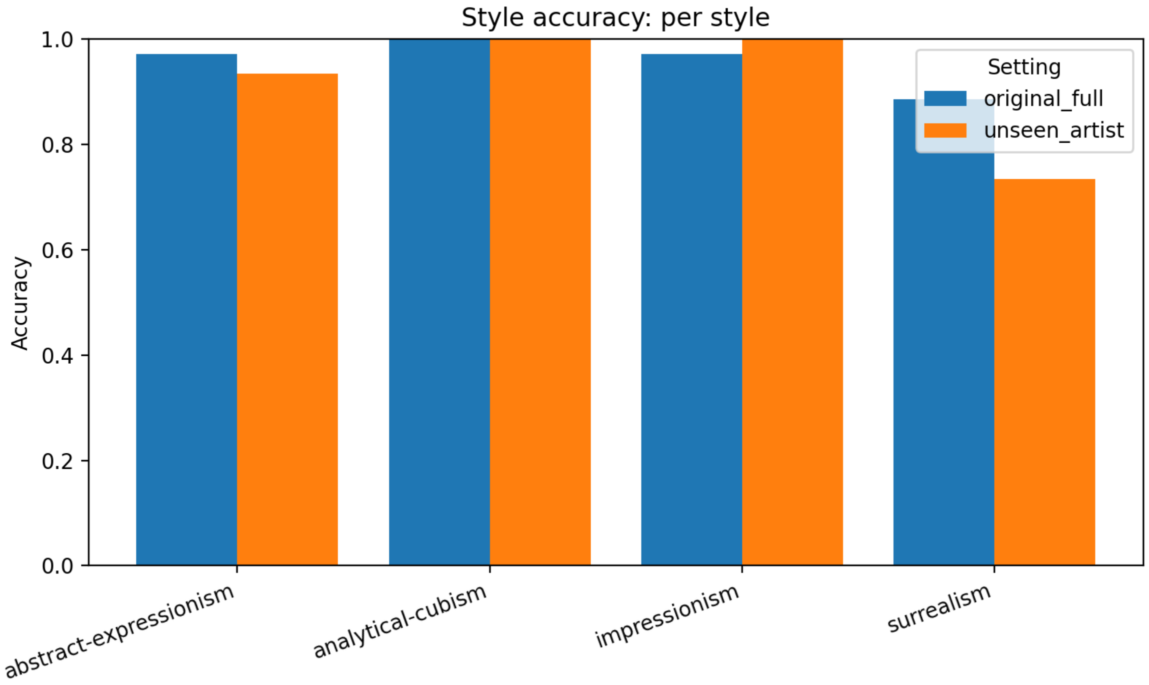

Figure 1

To reflect variability in the unseen-artist case, Table 3 adds the standard deviations from the 10 runs.

Style	Mean accuracy	Standard deviation
Abstract Expressionism	0.960	0.080
Analytical Cubism	1.000	0.000
Impressionism	1.000	0.000
Surrealism	0.700	0.184

Table 3: Unseen-artist per-style accuracy (10-run mean ± s.d.)

The main pattern can be summarised as follows:

- **Analytical Cubism** is classified perfectly in both settings. Accuracy remains 1.000 with zero variation across the 10 unseen-artist runs.
- **Impressionism** also reaches perfect mean accuracy in the unseen-artist evaluation. In these runs, every Impressionist query painting was assigned the correct style using only other artists as neighbours.
- **Abstract Expressionism** shows a small drop from 0.971 to a mean of 0.960, with some run-to-run variation (standard deviation 0.080).
- **Surrealism** is the most fragile style. Its accuracy falls from 0.886 in the original evaluation to a mean of 0.700 under the unseen-artist protocol, with a relatively large standard deviation (0.184).

3.3. Sample size and variability

Each unseen-artist run uses around 20 query paintings (one held-out artist per style), so individual misclassifications have a noticeable effect on the numbers. Repeating the experiment 10 times and averaging stabilises the estimates, although the standard deviations, especially for Surrealism, remain substantial.

For example, with five Surrealist query paintings in a run, each misclassification changes the Surrealism accuracy by 0.2. This explains the relatively large standard deviation (0.18) observed for Surrealism across runs. In contrast, Analytical Cubism and Impressionism happen to be consistently well separated in CLIP space for the artists in this dataset, leading to perfect accuracy and zero variance in the 10 runs.

Overall, the multi-run evaluation supports the earlier single-split results and makes the uncertainty around the numbers more explicit.

4. Discussion

The multi-run unseen-artist experiment adds extra context to the original style results in two main ways.

First, the modest drop in overall accuracy from 0.957 to about 0.922 (±0.035) shows that CLIP's image embeddings support style prediction beyond the artists seen as neighbours. For three of the four movements, mean unseen-artist accuracies remain very high. Analytical Cubism and Impressionism are effectively unchanged; Abstract Expressionism drops slightly but stays above 0.95 on average. This behaviour suggests that CLIP captures movement-level visual regularities that are shared across artists within these categories, such as characteristic compositions, colour palettes, and mark-making styles.

Second, the larger and more volatile drop for Surrealism reinforces the earlier finding that this movement is weakly clustered in CLIP space. In the original analysis, Surrealism already had the lowest k-NN accuracy and the least favourable silhouette scores. Under the unseen-artist protocol, its mean accuracy falls to 0.70 with relatively high variance. This suggests that CLIP's representation of Surrealism is less consistent. Some Surrealist artists and works align well with a distinctive cluster, while others lie closer to neighbouring movements and are easily reassigned to other styles once same-artist neighbours are removed.

At the same time, the limitations of the dataset remain important. The style subset is small (35 paintings per style, five artists per style) and not content-matched. Because subject matter, composition and medium differ systematically between styles, some of what CLIP seems to pick up as style is probably due to these factors rather than to style on its own in the art-historical sense. The unseen-artist protocol controls for artist identity but does not control for these broader content differences, so the results should be interpreted as evidence about how CLIP organises this particular dataset, not as a definitive statement about all twentieth-century painting.

Methodologically, this report shows how a simple change in evaluation protocol can reveal hidden assumptions in nearest-neighbour analyses. In the main report, k-NN accuracy was used primarily as a descriptive measure of clustering structure in the full dataset. By holding out artists from the neighbour pool and repeating the split across multiple random draws, the same embeddings and classifier now address a more demanding question about how far style-relevant features in CLIP generalise to new artists, and how stable that generalisation is across different choices of held-out artists.

In summary, the unseen-artist experiments support a nuanced view:

- CLIP's embedding space does contain meaningful information about art-historical style, especially for movements with strong, visually consistent signatures across artists.
- This information is not uniform. Some styles, such as Surrealism in this dataset, are represented more loosely and are more vulnerable to changes in which artists are available as neighbours.
- Evaluation design matters. Results that appear very strong when artists are mixed randomly between training and test sets can look more modest, and perhaps more realistic, when whole artists are held out from the evaluation.

These observations highlight the value of small, targeted follow-up experiments when interpreting the behaviour of large pretrained models in specialised domains such as art history.