Stochastic Interpolants for Revealing Stylistic Flows across the History of Art

Supplementary Material

A. Implementation details

Training. We set the number of transformer blocks $N = 12$ with a latent dimension of $2,048$. We used a batch size of $2,048$ and Adam optimizer [35] with a learning rate of $1e-4$ to train our model. The model is trained on a single A100 GPU for 20 hours. The backbone transformer of our model is depicted in Figure S8. All the style space share the same dimensionality of 768.

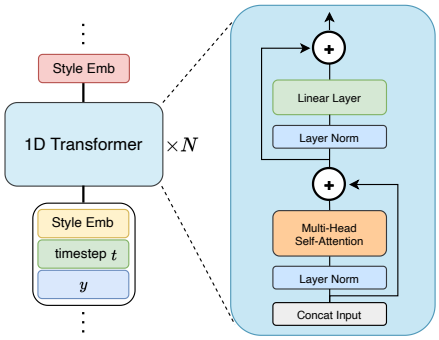

Figure S8. The backbone transformer. The condition consists of time step t and chronological information from the artwork y. These conditions are concatenated with the style embedding.

Using year as condition. This is a deliberate conceptual choice: instead of modeling stylistic evolution through static, human-defined categories, we treat it as a continuous phenomenon grounded in visual data. Categorical labels such as "Impressionism" are often coarse, overlapping, or inconsistent, especially in modern periods, and struggle to capture subtle stylistic shifts. More importantly, such discrete compartments limit our ability to trace how styles evolve, merge, or diverge over time. Treating style as a continuous function of year allows us to model artistic change as a fluid trajectory rather than as jumps between isolated labels.

While continuous conditioning introduces some ambiguity, this reflects the complexity of real-world art history, where multiple styles often coexist within the same period. For evaluation and reference, we include artist and style labels in the dataset, though they are not used during training or inference.

Style Representations. To systematically evaluate the suitability of different style representations, we considered four candidate feature spaces: CLIP [62], CSD-S, CSD-C [79], and DINOv2. We evaluated their effectiveness on our proposed dataset using standard retrieval metrics—mean Average Precision at top-k (mAP@k) [52] and Recall@k [31], following the evaluation protocol introduced in CSD [79].

As shown in Table S3, CSD-S achieves the highest performance in style retrieval, while CLIP performs slightly better in retrieving artworks by year. CSD-S, CSD-C, and CLIP all exhibit competitive performance. In contrast, DINOv2 lags behind, likely due to its emphasis on semantic content over stylistic variation.

In addition to retrieval metrics, we examined the structure of the embedding spaces via UMAP [48] visualizations, as shown in Figure S9. These projections highlight stylistic grouping behaviors across feature spaces: CSD-S embeddings show the most distinct structure, while CLIP and CSD-C maintain moderate separation. DINOv2 embeddings appear less organized but still retain some stylistic signal.

We also demonstrate an added benefit of using CLIP: it can be integrated with unCLIP [65] to generate plausible visual transformations of projected embeddings, as illustrated in Figure S15.

B. Limitations

Our study highlights a novel and complex task, using generative models to study how artistic styles flow across time, which presents several limitations.

Lack of well-established metrics. There is no established metric to evaluate this task, as the ground truth for stylistic correspondences across centuries is inherently absent. While we explored retrieval-based alignment and proposed measures as compactness δ and triplet consistency τ to assess consistency of temporally mapped clusters, these only partially address the underlying challenge.

Eurocentric dataset. Our proposed dataset primarily reflects Western artistic traditions, with a strong emphasis on *European artworks*. This curation choice was influenced by the domain expertise of our collaborators, ensuring historical accuracy and annotation quality. However, it introduces a cultural bias that may limit generalization to underrepresented artistic traditions. Addressing this imbalance is an important direction for future work.

Reliance on certain feature spaces. Although we trained and evaluated the model using four different style embeddings of varying quality, only the CLIP embedding currently supports rendering back to image space via the unCLIP

model [65]. While this enables qualitative inspection of temporal transformations, it limits our ability to visually analyze outputs from stronger stylistic representations such as CSD-S or CSD-C. Training unCLIP-like decoders for these spaces could offer more accurate visualizations in future iterations.

(a) CSD-C (b) CSD-S

(c) CLIP (d) DINOv2

Figure S9. 2-D UMAP for four candidate style spaces. We use the same subset of samples for all four plots but colored differently according to their ground truth style and year label. 800 samples were used for the top-8 prominent styles with a creation year coverage over 5 centuries.

C. Qualitative comparison

C.1. Ours vs. SD and FLUX

As mentioned in the main paper, when using *CLIP* as the style space, we can combine our model with a pre-trained Unclip model to render our generated CLIP embeddings (see Figure S15). Even with slightly worse performance on retrieval and mapping styles through time, this property makes it more attractive.

In the following, we present the reader some uncurated samples from individual centuries generated by *Stable Diffusion* and ours in Figure S10, Figure S12, and Figure S11. They are a subset of samples that we used to compute the metric in Table 2. For SD [66], SDXL [60], and FLUX [38] the text prompts are {"a professional artwork from year *xx*", "a professional painting created in *xx*", "An artwork from the *xx* th century"}. For our proposed method, we use random year numbers with the century as y to get the style embedding for the given year and feed it to the Unclip model to generate the samples for visualizations. We use the *Stable Unclip* model to transfer clip embeddings to RGB images.

Our model effectively captures the stylistic nuances over time, yielding visually appealing outcomes. This is also supported by the quantitative analysis of FID scores presented in the main paper (see Table 2). In contrast, alternative methods generate aesthetically pleasing artworks for the 16th, 17th, and 18th centuries but exhibit limited diversity across samples. It is important to note that they fail to encapsulate the evolving trends in art for later periods, producing only some old photographs of those times.

C.2. Matching Style Distributions

Besides using CLIP space with unCLIP to render samples, we also show that the other style embeddings produced by our method align well with the ground truth style distribution for different times. We show a few histogram plots for *CSD-S* embedding with only chronological condition y given in Figures S16 to S18. We observe a good match between the distributions formed by the generated embedding and the ground truth over styles, whose information was never provided to the model during training or inference. They are classified using k-NN and compared against the underlying distribution formed by ground-truth style labels. Figure S16 illustrates that our model closely replicates the ground-truth style distributions of the corresponding century when generating unconditionally. Similarly, Figures S17 and S18 shows that even when conditioning on existing style embeddings, the mapped distributions remain consistent with ground-truth labels, preserving recognizable hubs of retrieved styles.

D. Dataset details

The curated dataset proposed in the paper is under the most open *Creative Common CC-BY 4.0. The original images of artworks are under their copyright.*

The data was scraped from public websites using Selenium* and Beautiful Soup* in a similar way as LAION [74] (indexes to the internet in the form of URLs to the original image and meta information). The data was obtained from the following data sources:

- WikiArt.org, an online visual arts encyclopedia,
- Meisterdrucke, an art reproduction company with a public art gallery,
- Google Arts & Culture, an online art collection,
- Kaggle Best Artworks of All Time, a public dataset,
- Art UK, an online art collection,
- Tate, an institution that houses art galleries,
- The Museum of Modern Art (MoMA), an art museum,
- Web Gallery of Art, a virtual museum and searchable database of arts.

After obtaining the corresponding style embeddings, we subsequently discarded all the images of those paint-

*https://www.selenium.dev/
*https://www.crummy.com/software/BeautifulSoup/

	Style Recall@k ↑					Style mAP@k ↑			Year Recall@k ↑					Year mAP@k ↑		
Style space	1	2	5	10	100	5	10	100	1	2	5	10	100	5	10	100
DINOv2 base [55]	48.59	60.17	73.67	81.44	95.68	55.61	52.48	34.04	52.69	60.26	70.97	79.44	97.13	57.23	54.16	31.96
CSD-C ViT-L [79]	60.14	72.33	85.01	91.03	98.35	66.95	63.46	45.90	59.10	68.66	80.95	88.53	98.89	64.52	60.66	39.03
CLIP ViT-L [63]	59.53	72.19	85.09	91.01	98.61	66.69	63.21	45.85	**60.82**	**70.32**	**82.15**	**89.39**	**98.95**	**66.02**	**61.91**	**39.44**
CSD-S ViT-L [79]	**60.32**	**72.47**	**85.15**	**91.08**	**98.68**	**67.10**	**63.59**	**46.07**	59.07	68.75	81.12	88.69	98.93	64.54	60.65	38.96

Table S3. Style spaces evaluated by mAP and Recall metrics on our dataset. We omitted mAP@1 as it is equivalent to Recall@1.

	CSD-S										CLIP									
	SDEdit [49] Average Recall@k ↑					Ours Average Recall@k ↑					SDEdit Average Recall@k ↑					Ours Average Recall@k ↑				
Artistic Styles	1	5	10	$\delta\downarrow$	$\tau\uparrow(\%)$	1	5	10	$\delta\downarrow$	$\tau\uparrow(\%)$	1	5	10	$\delta\downarrow$	$\tau\uparrow(\%)$	1	5	10	$\delta\downarrow$	$\tau\uparrow(\%)$
Rococo	6.30	15.78	27.51	0.5716	45.6	**72.45**	**85.12**	**90.34**	**0.0235**	**84.8**	5.12	13.67	22.45	0.4623	44.8	**67.75**	**80.34**	**89.89**	**0.0182**	**70.2**
Impressionism	5.30	13.39	24.67	0.6051	34.5	**64.99**	**76.25**	**78.67**	**0.0241**	**75.8**	4.33	12.67	20.33	0.4687	30.4	**63.50**	**78.25**	**82.75**	**0.0183**	**75.5**
Realism	3.67	12.40	22.99	0.6125	45.6	**58.50**	**69.25**	**75.37**	**0.0253**	**68.6**	3.00	12.64	21.52	0.4715	38.9	**56.00**	**71.25**	**75.50**	**0.0184**	**67.4**
Romanticism	1.67	1.06	19.33	0.6305	37.8	**58.54**	**69.50**	**76.75**	**0.0297**	**58.4**	2.37	8.31	17.32	0.4842	41.2	**57.86**	**71.51**	**75.70**	**0.0185**	**59.6**
Post-Impressionism	2.17	7.68	16.44	0.6363	42.9	**67.33**	**82.23**	**88.00**	**0.0245**	**84.5**	2.74	9.60	14.80	0.4947	43.5	**66.25**	**80.75**	**87.57**	**0.0182**	**77.8**
Fauvism	1.24	4.53	11.06	0.3486	28.7	**55.67**	**65.89**	**70.23**	**0.0302**	**52.7**	1.56	5.34	9.12	0.5917	25.7	**55.34**	**70.12**	**74.56**	**0.0190**	**47.7**
Expressionism*	2.20	5.38	12.65	0.7415	36.2	**70.33**	**82.33**	**86.67**	**0.0252**	**74.2**	1.09	6.33	10.47	0.5824	34.6	**58.66**	**75.67**	**81.38**	**0.0189**	**65.9**
Surrealism*	1.80	3.49	10.54	0.7943	23.4	**50.78**	**60.45**	**68.56**	**0.0315**	**46.3**	1.23	4.78	8.34	0.6025	22.3	**54.78**	**69.45**	**73.89**	**0.0192**	**41.3**

Table S4. Quality of the time matching when mapping artworks through time The mean value of multiple time jumps is calculated for a year range of $[-100, 100]$ with 25 step size. We sampled a compact subset of a hundred samples for each style to calculate the metrics. For τ, we used the $nns[:25]$ neighbors as positive samples and $nns[50:75]$ as the negative samples for each chosen anchor, where nns denotes its nearest neighbors. Underscore denotes the best value column-wise. *: *for these styles, we only perform jumps up to* $+50$ *as a larger jump will result in a region without any ground-truth samples.*

ings/artworks. Any researcher using the datasets must reconstruct the image data by downloading the subset they are interested in. We maintain a list of valid URLs to the original image as part of the meta-information for each sample. As artworks were obtained from various data sources and the information accompanying them varies, not all images come with the complete set of metadata.

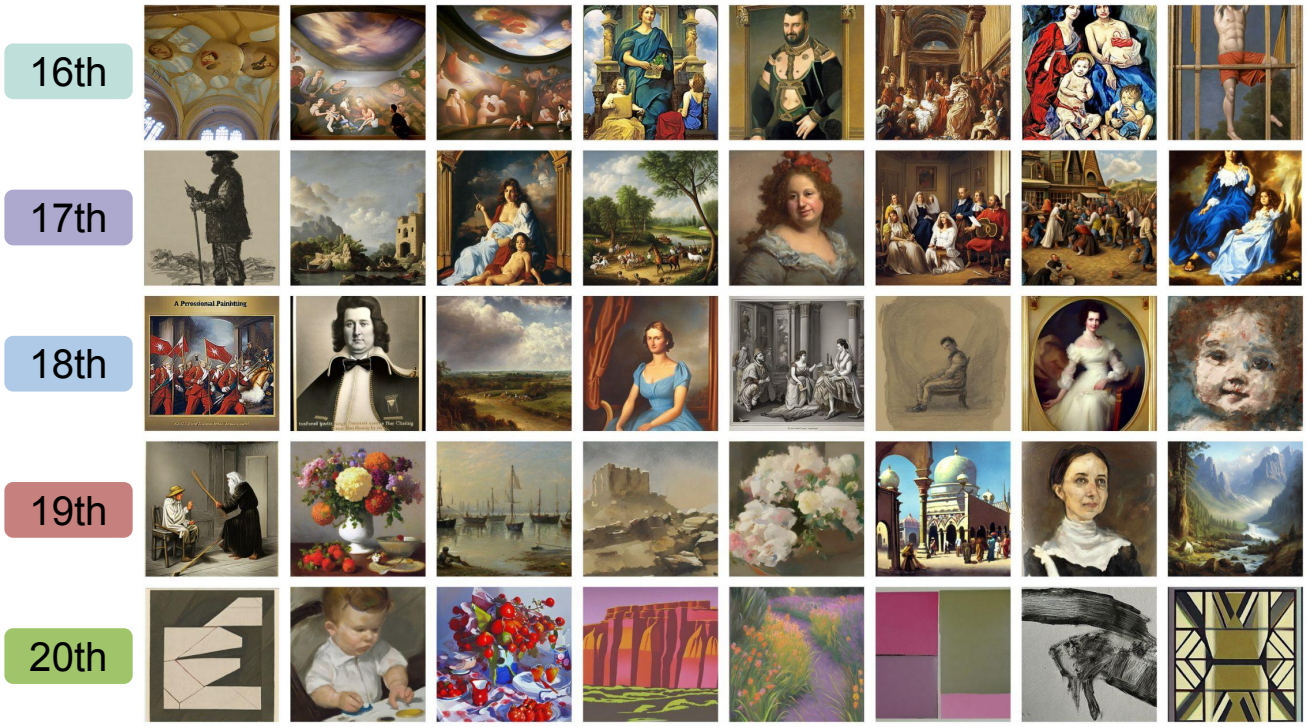

Figure S10. Uncurated samples generated by our method. Each row represents a collection of unconditionally generated artworks from a specific century.

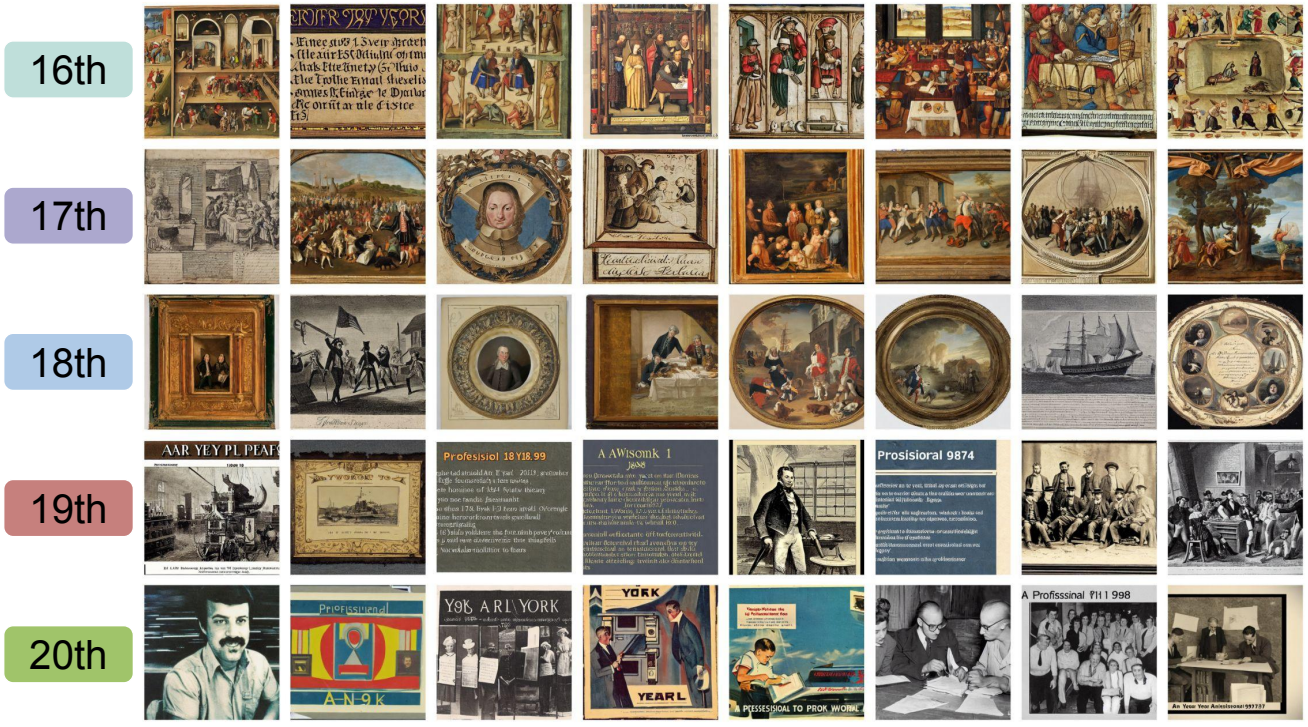

Figure S11. Uncurated samples generated by SD1.5. Each row represents a collection of randomly generated artworks from a specific century.

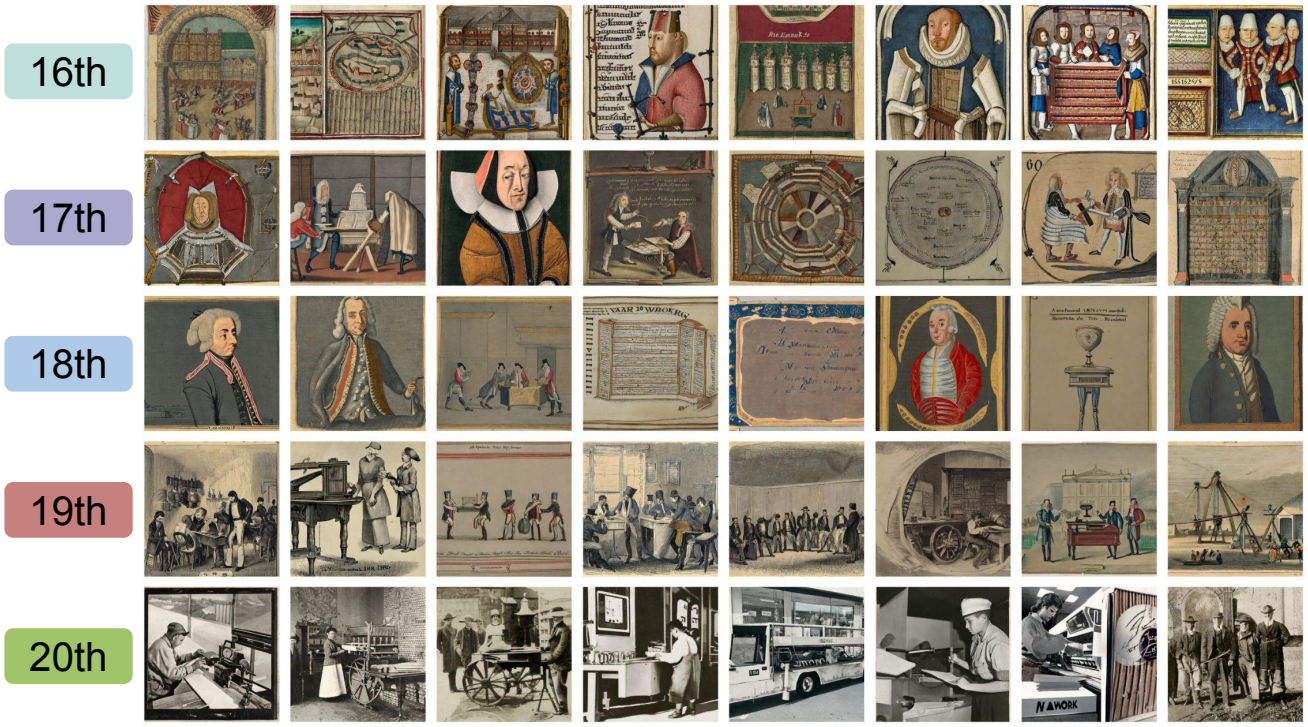

Figure S12. Uncurated samples generated by SD2.1. Each row represents a collection of randomly generated artworks from a specific century.

Figure S13. Uncurated samples of SDXL. Each row represents a collection of randomly generated artworks from a specific century.

Figure S14. Uncurated samples of FLUX.1-schnell. Each row represents a collection of randomly generated artworks from a specific century. Albeit its better image quality and details due to its model size and much stronger backbones, FLUX still fails to capture the underlying artistic style distribution, resulting in a limited variety of styles.

Figure S15. Generated samples from Unclip [65, 66]. The styles are well aligned with the reference image. The large image on the left is the reference image. The corresponding four small images are Unclip results from the same reference image.

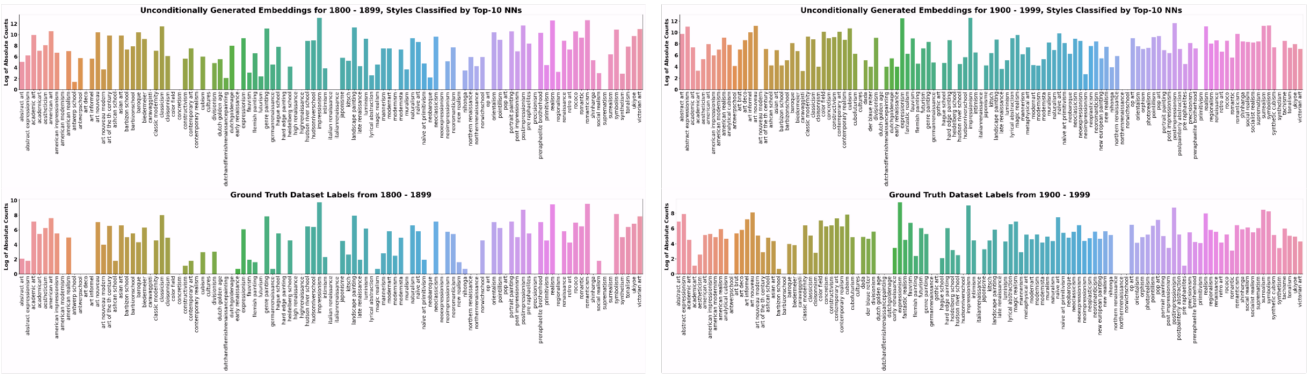

Figure S16. Discrete style distributions of unconditionally generated embeddings (CSD-S) for different periods classified by nearest-neighbors. For each generated sample, style labels are predicted by a 10-nn classifier (upper row). Log-scale of absolute count is used to make small values more visible. The bottom row shows the corresponding distribution of ground truth labels. This denotes our model's ability to match the artistic style distortion with the given century. The x-axis depicts the aligned style labels.

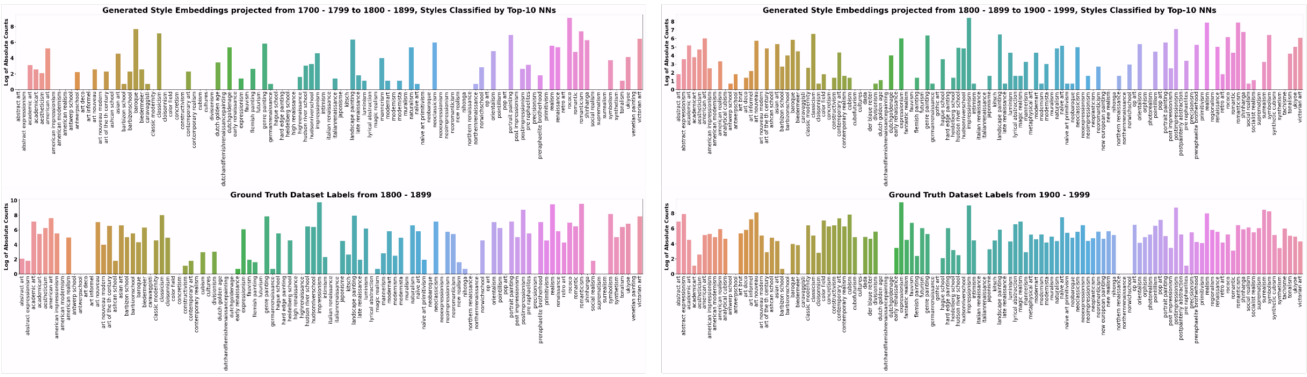

Figure S17. Discrete style distributions of artworks transformed from 18th to 19th (left) and 19th to 20th (right) century. Distributions are retrieved in the same way as in Figure S16 but the conditioning is based on style embeddings from an earlier century instead of Gaussian noise. Similar hubs of retrieved styles are visible.

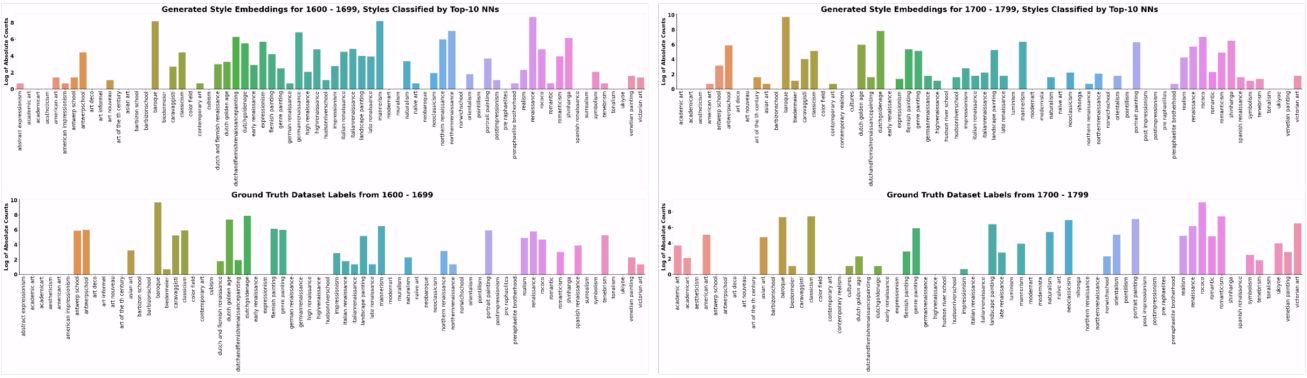

Figure S18. Discrete style distributions of artworks transformed from 16th to 17th (left) and 17th to 18th(right) century. Style Distributions for further centuries as in Figure S17. Similar hubs of retrieved styles are visible.