IntroStyle: Training-Free Introspective Style Attribution using Diffusion Features

Anand Kumar Jiteng Mu Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
{ank029, jmu, nvasconcelos}@ucsd.edu

Abstract

Text-to-image (T2I) models have gained widespread adoption among content creators and the general public. However, this has sparked significant concerns regarding data privacy and copyright infringement among artists. Consequently, there is an increasing demand for T2I models to incorporate mechanisms that prevent the generation of specific artistic styles, thereby safeguarding intellectual property rights. Existing methods for style extraction typically necessitate the collection of custom datasets and the training of specialized models. This, however, is resource-intensive, time-consuming, and often impractical for real-time applications. Moreover, it may not adequately address the dynamic nature of artistic styles and the rapidly evolving landscape of digital art. We present a novel, training-free framework to solve the style attribution problem, using the features produced by a diffusion model alone, without any external modules or retraining. This is denoted as introspective style attribution (IntroStyle) and demonstrates superior performance to state-of-the-art models for style retrieval. We also introduce a synthetic dataset of Style Hacks (SHacks) to isolate artistic style and evaluate fine-grained style attribution performance.

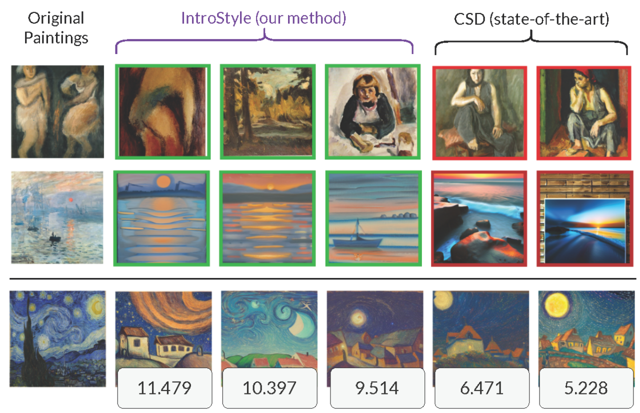


Figure 1. Introspective Style Attribution (IntroStyle). Existing SOTA methods, e.g., CSD, are biased to content semantics and fail to retrieve images of similar styles, underscoring the need for a better style similarity metric. The top two rows show examples on the WikiArt [33] dataset and our proposed synthetic Style Hacks (SHacks) dataset, respectively, where a reference image, top-3 retrieval results of our method, top-2 retrieval results of CSD are shown from left to right. More importantly, our proposed method can be used as a metric for style measurement, as shown in the third row, with a lower score indicating images further away in style from the reference in the first column. Green colors indicate correct and red for incorrect retrievals.

1. Introduction

Diffusion models have significantly advanced image synthesis by applying iterative denoising processes guided by input prompts. Models like Stable Diffusion [30], DALL-E 3 [1], and Imagen [32] have emerged as a powerful paradigm for text-to-image synthesis, demonstrating remarkable capabilities in generating high-quality images from textual descriptions. The success of diffusion-based approaches has led to their application in various domains, including layout-to-image generation, text-guided image generation, and even video synthesis. However, these models' widespread adoption and impressive performance have also raised concerns about problems like copyright infringement. For effective performance, the models require large-scale pre-training on diverse datasets [34]. Since these datasets are collected automatically and dominantly from the web, it is difficult to control image provenience and avoid the collection of copyrighted imagery. This problem is compounded by the tendency of diffusion models to replicate elements from their training data [2]. The legal implications of training models on copyrighted images have become a subject of recent debate and litigation, with artists arguing that the unauthorized use of their works for AI training [7, 10, 12] constitutes copyright infringement [17, 29].

Current mitigation approaches include "unlearning" techniques to remove specific styles from AI models [44], though these require costly model retraining and may not fully address indirect style replication through alternative prompts [24]. Style "cloaking" methods [35] offer artists some protection but compromise the authentic viewing experience and place the burden of protection on creators. An alternative solution is to rely on attribution techniques [37, 43, 44], which can be used to attribute synthesized images to different artistic styles, allowing the post-hoc assessment of how much they might, for example, replicate an artist's artwork. These methods are also more practical since they simply return examples from the dataset used to train the model, ranked by similarity to the synthesized image.

Prior attribution approaches have proposed retraining the diffusion model [44] or adding external models to evaluate data similarity [37, 43]. This can add significant complexity since effective similarity models, like CLIP [26] or DINO [3], can be comparable in size to the diffusion model. On the other hand, prior diffusion model research has shown that different diffusion model layers develop distinct representations of higher and lower level image properties, such as structure vs color [42], producing features that can be leveraged to solve problems like image correspondence [39] or segmentation [40] without retraining. This led us to hypothesize that it may be possible to solve the attribution problem using the features produced by a diffusion model alone, without any external modules or retraining. We refer to this as *introspective style attribution* (`IntroStyle`).

In this work, we show that the hypothesis holds. We propose a simple and effective approach to differentiate image styles by analyzing the similarity of the features produced by a diffusion model for different images. We frame the denoising network of the diffusion model as an autoencoder whose encoder produces stylistic features. We then propose a metric to compare these features that, while quite simple, are shown to be quite effective for style retrieval. In fact, a comparison with the current state-of-the-art on popular style retrieval datasets shows that `IntroStyle` significantly outperforms prior models trained in the existing evaluation datasets, as illustrated in the top rows of Fig. 1. Though the existing models are fine-tuned for style similarity, they still have semantic bias, whereas our method focuses solely on style.

Beyond models, the problem of style attribution suffers from a shortage of datasets for evaluation. Creating such datasets is far from trivial since even the concept of "style" in art remains debatable. Nevertheless, many recognized artistic styles, such as surrealism and impressionism, are often associated with specific artists or movements. We use this social construct to define style as the collective global characteristics of an image identifiable with a particular artist or artistic movement. These characteristics encompass various elements such as color usage, brushstroke techniques, composition, and perspective. Under this definition, a single artist can produce art in several styles. Current datasets inadequately address AI-generated copyright concerns, primarily focusing on comparisons between authentic artworks rather than evaluating against AI-synthesized images. While CSD [37] introduces synthetic data for generated image evaluation and GDA[43] uses custom diffusion to generate training datasets, these existing datasets fail to account for prompt manipulation techniques that can circumvent copyright protections [4, 41, 45].

To address these problems, we create the *Style Hacks* (`SHacks`) dataset with inspiration from [4]. This dataset is based on real image queries that rank synthetic images created by manipulating their prompts. For each image, we produce a set of "innocent" prompts, which do not mention style, and a set of "hacked" prompts, which indirectly mention style in a way that is not easy to detect by text analysis. These prompts are then fed to a diffusion model to create corresponding images. By measuring similarities between features extracted from the query and synthetic images, it is then possible to perform fine-grained evaluations of the suitability of the feature representation for style attribution. Hence, style retrieval experiments on this dataset help evaluate the capability of different models to address problems like copyright infringement.

The second row of Fig. 1 shows retrieval results on the *Style Hacks* (`SHacks`) dataset, where the proposed model outperforms a state-of-the-art approach. Furthermore, the proposed style attribution approach allows improved sorting of images by similarity to a reference. An example is shown in the last row of Fig. 1, showing how images of style closer to the reference have higher similarity scores. This can be leveraged for style-based rejection to prevent diffusion models from generating images similar to a specific style.

Overall, this paper makes the following contributions:

1. We formulate the problem of introspective style attribution without model retraining or external modules.
2. We propose `IntroStyle`, a method to generate style descriptors from images and a metric to compare image pairs according to style.
3. We introduce the `SHacks` dataset to isolate artistic style and evaluate fine-grained style attribution performance.
4. We perform extensive experiments showing that `IntroStyle` outperforms existing approaches of much higher complexity.

2. Related Work

2.1. Style Transfer

Style attribution and retrieval are related to style transfer, where the goal is to render one image in the style of another. Early approaches include [49], which formulated style transfer using Markov random fields. However, the topic raised in popularity with the rise of deep learning and the work of [9], which introduced Gram Matrices as style descriptors and an optimization framework for style transfer. Many variants have been proposed, including various regularizations to prevent distortions in reconstructed images [20] or representations of correlation beyond Gram matrices [5]. However, recent attribution work shows that Gram-style representations cannot guarantee effective style attribution [37].

Our work is instead inspired by more recent advances in style transfer using Generative Adversarial Networks (GANs) and other generative models, where methods like Adaptive Instance Normalization (AdaIN), which aligns the mean and variance of content features with those of style features, have been shown beneficial for real-time style transfer [14]. These ideas are central to the StyleGAN [15], where AdaIN layers are embedded into the generator architecture. This allows fine-grained control over different aspects of the generated images, from coarse features like pose to fine details like color schemes.

Recent advances in diffusion-based style transfer demonstrate promising results, notably through AdaIN layer manipulation of initial timestep latent [6], and DDIM Inversion-based style latent reuse [21]. These approaches informed our investigation into diffusion model latent distributions.

2.2. Style-aware Text-to-Image Models

Early work in style-aware text-to-image generation built upon the success of neural style transfer techniques. Reed et al. [28] introduced the first GAN-based text-to-image model, while [48] improved image quality and resolution through a stacked architecture. More recently, diffusion models like DALL-E 2 [27] and Stable Diffusion [30] have demonstrated impressive capabilities in generating stylized images from text prompts. Several approaches have focused on explicitly representing and manipulating style in text-to-image models, such as style-based GAN architectures with fine-grained style control [19] or textual inversion methods [8, 16, 21] to learn new concepts and styles from a few examples, enabling personalized text-to-image generation. All these works require some fine-tuning of the generative model, whereas we pursue training-free solutions to the style attribution problem.

2.3. Data Attribution

While there is extensive literature on image retrieval, most methods focus on semantic similarity, i.e., the retrieval of images of similar content in terms of objects, scenes, etc. Style is a subtle property not well captured by representations, such as DINO [3] or CLIP [26], developed for this purpose. Few techniques have been developed specifically for style retrieval or attribution. Lee et al. [18] employed two separate neural network modules to facilitate style retrieval for image style and content. Recent works [31, 37, 43] instead train attribution models using contrastive objectives based on either Behance Artistic Media (BAM) dataset [31], a subset LAION Aesthetic Dataset [34] or synthetic style pairs. [31] proposes a dual encoder to separate content and style with style features obtained using AdaIN layers, whereas CSD[37] and GDA[43] fine-tune a pre-trained CLIP or DINO backbone with a linear layer on top with the former training the entire network and latter training the linear layer alone. It is also possible to formulate data attribution as the unlearning of target images [44], but this requires a different model for each target. Unlike these methods, we seek an introspective solution to style attribution that does not require external models or attribution-specific training.

2.4. Diffusion Features

In the diffusion model literature, various works have studied the representations learned for denoising, showing that different layers of the network learn features that capture different image properties, like structure, color, or texture [36, 42, 50]. Denoising Diffusion Autoencoders (DDAE) [46] leverage intermediate activations from pre-trained diffusion models for downstream tasks, demonstrating effectiveness in image classification through linear probing and fine-tuning. Recent advances include REPresentation Alignment (REPA) [47], which enhances semantic feature learning by aligning diffusion transformer representations with self-supervised visual features. Additionally, diffusion model features have shown promise in zero-shot applications like correspondence [39] and segmentation [40]. These observations have inspired our hypothesis that style attribution can be solved introspectively.

3. Method

3.1. Diffusion

Diffusion models [38] are probabilistic models that define a Markov chain from which images are sampled by gradually denoising a normally distributed seed, using a neural network ϵ_θ. T2I models condition this sampling on a text prompt. A sentence describing the picture is encoded into a text embedding c by a text encoder and then used to condition the denoising network, usually via cross-attention. Stable Diffusion [30], a prominent T2I model, learns the chain in a latent space of a pre-trained image autoencoder. An encoder $\mathcal{E}$ first maps an image into a latent code z_0. In a forward process, a sequence of noisier codes z_t are obtained by adding Gaussian noise to z_0 according to

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t, \quad \epsilon_t \sim N(0, I) \tag{1}$$

where $t \in \{0, 1, \dots, T\}$ is a timestep, $T = 1000$, $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$ and α_i are pre-defined according to a noise sched-

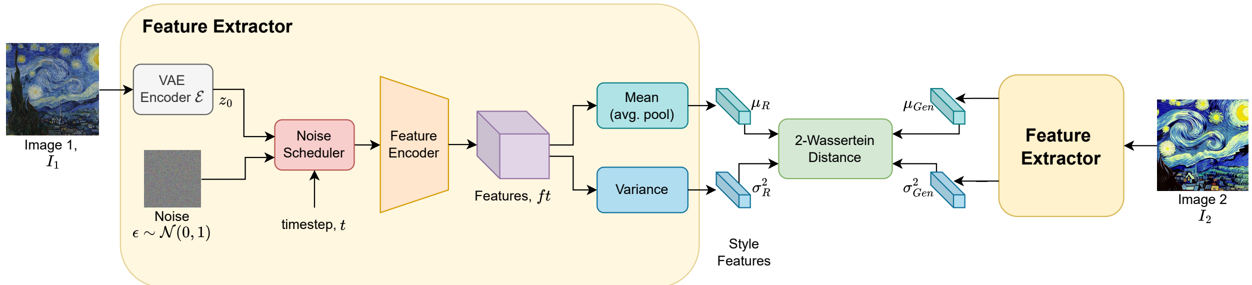


Figure 2. `IntroStyle` computation of style similarity: channel-wise mean μ and variance σ^2 are computed for the identified style layers. Then a distance metric, 2-Wassertein Distance, can be used to measure styles between a pair of images.

ule. In a reverse process, the noise latent z_T is used as the seed for a denoising chain, based on the network ϵ_θ, which aims to recover the noiseless latent z_0. This is finally fed to a decoder $\mathcal{D}$ to recover the image. The network parameters θ are trained to predict the noise ϵ_t introduced at each step of the forward process, using the loss

$$L_{SD} = \mathbb{E}_{z,\epsilon_t \sim N(0,I),t} \|\epsilon_t - \epsilon_\theta(z_t, t, c)\|_2^2. \tag{2}$$

During inference, the reverse process generates a clean latent $\hat{z}_0$, starting from a random seed $z_T \sim N(0,1)$, conditioned on a text embedding c. For stable diffusion, the model ϵ_θ has a UNet architecture with 4 downsample, 1 mid, and 4 upsample blocks, as shown in Fig. 3.

3.2. Style attribution

Style attribution aims to identify in a reference dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$ of images x_i, usually but not necessarily the dataset used to train the diffusion model, the subset of images of style closest to that of a query image q, which is usually an image synthesized by the model. Like all retrieval operations, this requires two components: 1) a feature representation $\mathcal{F} : x \rightarrow f$ that maps each image x into a feature tensor f that captures the stylistic elements of x, and 2) a similarity metric $d(\mathcal{F}(q), \mathcal{F}(x))$ that quantifies the style similarity of the feature tensors extracted from query q and reference x images. This allows retrieving the images in $\mathcal{D}$ that are most similar to q, which can then be inspected, e.g. to determine copyright violations.

3.3. Introspective style attribution

Style attribution is not a widely studied problem in the literature. Previous approaches [37, 43] attempted to train an external feature representation $\mathcal{F}$ or adapt existing feature representations, such as DINO [26], by addition of parameters specific to the extraction of style features. This has two major drawbacks: it requires computationally intensive large-scale training to cover the diffusion model's sampling distribution, and the external feature extractor $\mathcal{F}$ introduces substantial memory and computational overhead—often exceeding the diffusion model's resource requirements.

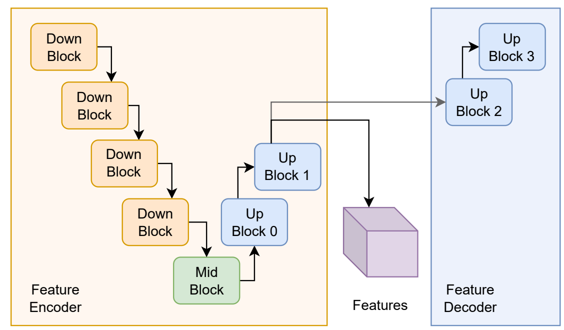


Figure 3. Features from the first up-block are used to extract style features. Stable Diffusion 2.1 is used in our experiments.

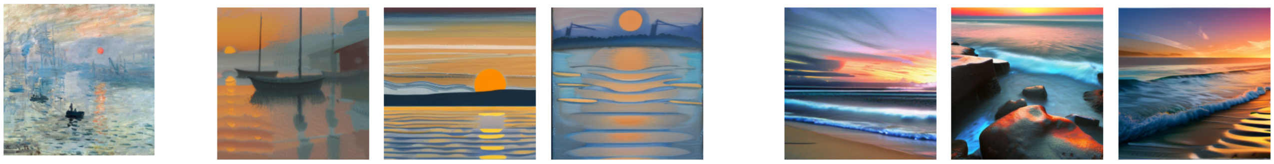
Introspective style attribution aims to solve the problem without external models by using only the features computed by the denoising network ϵ_θ of the diffusion model. This is inspired by recent showings that this network produces feature representations that 1) disentangle image properties such as structure and color [42, 50] and 2) are sufficiently informative to solve classical vision problems such as classification [46], correspondence [39], or segmentation [40] with good performance. This suggests that these features should also be sufficient for style attribution. In what follows, we show that this is indeed the case and can be achieved with very simple feature representations and popular similarity metrics, resulting in the simple `IntroStyle` approach, based on the operations of Fig. 2. The features and similarity metric used by `IntroStyle` are discussed next.

`IntroStyle` Features To compute a feature representation, image I is first mapped into a latent vector $z_0 = \mathcal{E}(I)$, which is noised to a time t using Eq. 1. The noised latent z_t is then passed through the denoising network ϵ_θ with null text embedding $c = \varnothing$. A feature tensor F^t is then extracted from intermediate network layers. Given the UNet architecture commonly used to implement ϵ_θ, we consider this model an autoencoder that stacks a sequence of feature encoder and decoder blocks. The encoder/decoder split is quantified by equating all layers up to the upsampling block

(a.1) Pablo Picasso's *Girl before a Mirror*

(a.2) *Create a Cubist-style portrait of a woman facing her reflection in a mirror. Use bold, contrasting colors and geometric shapes to depict both the figure and her distorted image.*

(a.3) *Create a portrait of a woman facing her reflection in a mirror.*



(b.1) Claude Monet's *Impression, Sunrise*

(b.2) *Create an Impressionist-style seascape at dawn, featuring a hazy orange sun reflecting on rippling water. Include silhouettes of boats and misty harbor structures in muted blues and grays.*

(b.3) *Create a seascape at dawn.*

Figure 4. *Style Hacks* (`SHacks`) dataset samples. Each row shows images generated for a reference image by an artist. Reference image, three images generated by "hacked" prompts, and three images generated by "innocent" prompts are shown from left to right.

idx as an encoder and the remaining upsampling layers as a decoder. The feature tensor $F^{t,idx}$ is finally extracted from block idx, as illustrated in Fig. 3 for $idx = 1$.

To derive a compact feature representation that captures style information, we then take inspiration from the literature on style-aware generative models and the successful StyleGAN approach to control style by Adaptive Instance Normalization (AdaIN) of first and second-order feature statistics [14]. This leads us to hypothesize that diffusion models may be intrinsically producing similarly normalized features, in which case these statistics should be highly informative of style and eliminate dependence on pixel-wise correlations.

Based on this intuition, we propose a simple style feature representation for `IntroStyle`, based on the channel-wise mean and variance of feature responses. Given the feature tensor $F^{t,idx}$, the mean and variance of channel c are obtained with

$$\mu_c = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} F^{t,idx}_{c,i,j}\,,\ \sigma_c^2 = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}(F^{t,idx}_{c,i,j} - \mu_c)^2$$

where W, H are the spatial dimensions of the tensor. The style feature representation of the image I is finally

$$f^{t,idx}(I) = (\mu_1, \ldots, \mu_C, \sigma_1^2, \ldots, \sigma_C^2)^T \qquad (3)$$

where C is the number of channels at block idx. Both t and idx are hyperparameters of the method, which is summarized in Figures 2 and 3.

Similarity Metric The `IntroStyle` representation of an image models its feature statistics with a multivariate Gaussian distribution of C dimensions. As shown in Figure 2, Gaussian parameters (μ_1, Σ_1) and (μ_2, Σ_2) are extracted for images I_1 and I_2, respectively. Hence, any similarity measure between two multivariate Gaussians can be used to compare the images. We have investigated several similarity measures and found the 2-Wasserstein (W_2) distance

$$\begin{aligned} W_2((\mu_1, \Sigma_1); (\mu_2, \Sigma_2))^2 &= \|\mu_1 - \mu_2\|_2^2 \\ &+ \operatorname{tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right). \end{aligned} \qquad (4)$$

to be effective for style retrieval. A comparison to other metrics, including Euclidean (L_2) Distance, Gram Matrix and Jensen-Shannon Divergence (JSD) is presented in the experiments. In all cases, the covariance matrices are assumed diagonal with the variances of the `IntroStyle` representation as diagonal entries.

4. The `SHacks` Dataset

To address limitations in existing attribution datasets, we introduce the *Style Hacks* (`SHacks`) dataset for evaluating style hacking detection. We selected prompt-image pairs for 25 prominent artists from the LAION Aesthetic Dataset [34], focusing on their most recognized works as queries. We used Stable Diffusion v2.1 for each artist to synthesize a reference dataset containing images produced in response to two types of prompts:"innocent" and "hacked". A set of such prompts was first generated for each prompt in the query set. "Innocent" prompts solely include a description of the image in terms of objects, scenes, etc., without mention of any style. "Hacked" prompts indicate the artist's style indirectly, without explicitly referencing an artist name or painting[1]. Both "hacked" and "innocent" prompts were generated from the original prompts using ChatGPT[22] and manually inspected to remove any mention of the artist or their paintings directly. We generated 3

[1] It is assumed that if the user directly mentions the artist, a simple text filter will suffice to prevent generation.

(a) Queries (b) `IntroStyle` (ours) (c) CSD [37] (d) GDA [43]

Figure 5. Image Retrieval on WikiArt Dataset for `IntroStyle`, CSD, and GDA. The first column (left to right) shows the query image, followed by `IntroStyle`'s top 3 retrieval results, CSD and GDA, respectively. The colored boxes convey retrieval groundtruth similar to Fig. 1

	WikiArt						DomainNet					
	mAP@k			Recall@k			mAP@k			Recall@k		
Method	1	10	100	1	10	100	1	10	100	1	10	100
VGG-Net Gram [9]	0.259	0.194	0.114	0.259	0.527	0.804	-	-	-	-	-	-
SSCD ResNet-50 [25]	0.332	0.243	0.135	0.332	0.591	0.839	0.713	0.672	0.608	0.713	0.964	0.998
MoCo ViT-B/16 [11]	0.440	0.332	0.188	0.440	0.689	0.879	0.784	0.757	0.708	0.784	0.960	0.997
DINO ViT-B/16 [3]	0.458	0.348	0.197	0.458	0.706	0.889	0.774	0.741	0.684	0.774	0.956	0.997
CLIP ViT-L [26]	0.595	0.489	0.316	0.595	0.830	0.952	0.928	0.923	0.778	0.928	0.972	0.996
GDA ViT-B [43]	0.439	0.331	0.186	0.439	0.685	0.873	0.800	0.768	0.694	0.800	0.966	0.998
GDA CLIP ViT-B [43]	0.522	0.419	0.259	0.522	0.782	0.935	0.817	0.786	0.716	0.817	0.972	0.998
GDA DINO ViT-B [43]	0.458	0.349	0.199	0.458	0.709	0.8888	0.770	0.739	0.684	0.770	0.956	0.997
CSD ViT-L [37]	0.646	0.538	0.357	0.646	0.857	0.956	0.833	0.811	0.766	0.833	0.972	0.998
`IntroStyle` (Ours)	0.887	0.852	0.525	0.887	0.909	0.941	0.954	0.949	0.850	0.954	0.982	0.995

Table 1. Evaluation Results on WikiArt and DomainNet datasets. `IntroStyle` outperforms all previous methods significantly. The best values are highlighted with red and second best with orange.

prompts of each of the two types per original prompt and used the diffusion model to generate 12 images per prompt. This produced a reference dataset with a total of 1800 images. Samples of images generated using this method are shown in Fig. 14.

The reference images are ranked by similarity to each query image to evaluate style attribution performance. Ideally, this would rank all the "hacked" prompt images above the "innocent" ones. Hence, a success is declared when a "hacked" image is retrieved. Standard precision and recall metrics then evaluate style attribution performance.

5. Experiments

We use the Stable Diffusion v2.1 model initialized with weights from Hugging Face. Since introspective attribution is training-free, we extract feature tensors from this model to obtain the `IntroStyle` representations for a given image, using an NVIDIA RTX 3090ti GPU with a batch size of 4. Each image is center-cropped to a resolution of 512×512, and the guidance scale does not matter as we don't have a text prompt.

5.1. Datasets and Baselines

Dataset. *DomainNet* [23] comprises images from six domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch, with approximately equal representation. Due to strong stylistic similarities between Quickdraw and Sketch domains, we excluded the former. The test set, consisting of $127,855$ images, was randomly divided into $25,571$ query images and $102,284$ reference images. *WikiArt* [33] contains $80,096$ fine art images representing $1,119$ artists across 27 genres. We randomly split this dataset into $64,090$ reference images and $16,006$ query images. As proposed in [37], we define each artist as a separate class, correlating style and artist. The WikiArt dataset is more challenging for style retrieval tasks, as the random chance of successful matching is 0.09%, significantly less than the 20% of DomainNet, due to a large number of artists. These are the only two publicly available datasets for style retrieval[2], making our `SHacks` dataset a valuable addition to the field. We also conduct experiments on `SHacks`, which

[2]The Behance Artistic Media (BAM) dataset [31] is not publicly available.

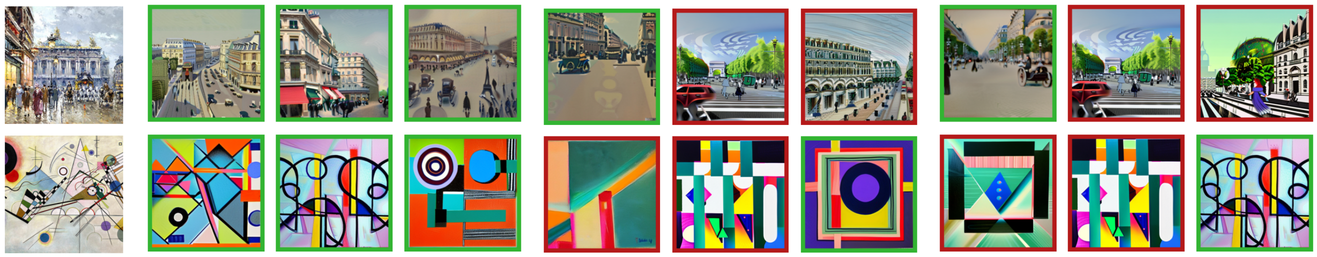


Figure 6. Image retrieval visualization on SHacks Dataset for IntroStyle, CSD, and CLIP. The first column (left to right) shows the query image, followed by IntroStyle's top 3 retrieval results, CSD and CLIP, respectively. The query images are based on the styles of Antoine Blanchard and Wassily Kandinsky. The colored boxes convey retrieval groundtruth, as in Fig. 1.

Method	mAP@5	mAP@10
DINO ViT-B/16 [3]	0.526	0.626
CLIP ViT-L [26]	0.794	0.782
GDA DINO ViT-B [43]	0.586	0.694
CSD ViT-L [37]	0.700	0.670
IntroStyle (Ours)	0.823	0.793

Table 2. Evaluation results on the SHacks dataset.

has 25 queries, each ranking 72 images.

Evaluation Metrics. Style attribution performance is measured by two standard retrieval metrics [37, 43]: Recall@k and mean Average Precision (mAP)@k, for $k \in 1, 10, 100$.

Baselines. We compared IntroStyle to several baselines from the style attribution and representation learning literature. For style attribution, we consider two recent methods, Contrastive Style Descriptors (CSD) [37] and Generative Data Attribution (GDA) [43][3], which employ fine-tuning on additional datasets for enhanced performance. This is complemented by a training-free method based on VGG Gram Matrices [9]. We utilize the final layer embeddings as feature representations for all models under comparison. This standardized approach ensures a fair comparison across diverse architectures and training paradigms, allowing us to isolate and evaluate the effectiveness of each representation for style attribution. For representation learning, we consider multiple vision foundational models, including CLIP [26], DINO [3], MoCo [11], and SSCD [25], which are proven effective across various computer vision tasks.

5.2. WikiArt and DomainNet Results

Table 1 shows that IntroStyle achieves the overall best performance, when compared to various baselines. We compare the attribution performance on the WikiArt and DomainNet datasets. IntroStyle is implemented with $idx = 1$, timestep $t = 25$, and the 2-Wassertein distance. Its gains are quite significant for the more practically relevant (lower) values of k. In the more challenging WikiArt dataset, IntroStyle achieves a gain of close to 20 points over the previous best approach (CSD) for both mAP and Recall when $k = 1$. For $k = 10$, the mAP gain is more than 30 points. The improvement drops at $k = 100$ because most artists have less than 100 images, making it impossible to achieve a perfect mAP score and masking the differences between methods. In DomainNet, when compared to style attribution approaches, IntroStyle has gains of around 10% for $k = 1$ in both the metrics and a similar improvement on mAP@10. We note that these are gains of large magnitude, not commonly seen in retrieval applications, suggesting that introspective style attribution is vastly superior to external models. Fig. 5 shows examples of improved retrieval.

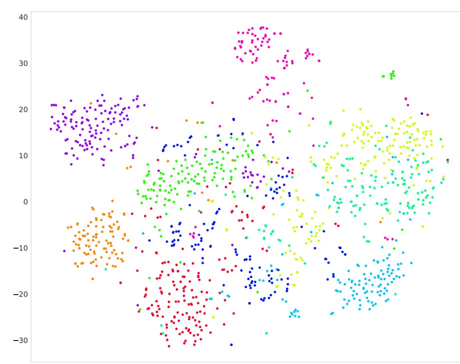


Figure 7. TSNE Clustering visualization on WikiArt images.

To better understand the reason behind these gains, we visualize how the IntroStyle representation clusters images of different artists. For this, we randomly choose 9 artists from WikiArt, each with 100 to 200 images. We represent each image by the mean component μ of its IntroStyle representation and visualize the TSNE clustering [13], shown in Fig. 11. Images of the different artists are colored differently. Results show the emerged style clusters, demonstrating the effectiveness of the extracted features. There are, however, different subclusters

[3] We use the real mapper to obtain features

Block	mAP@k 1	10	100	Recall@k 1	10	100
$idx = 0$	0.866	0.830	0.474	0.866	0.879	0.911
$idx = 1$	0.887	0.852	0.525	0.887	0.909	0.941
$idx = 2$	0.888	0.852	0.521	0.888	0.907	0.938
$idx = 3$	0.872	0.836	0.491	0.872	0.886	0.916

Table 3. Ablation on block index selection, all with $t = 25$.

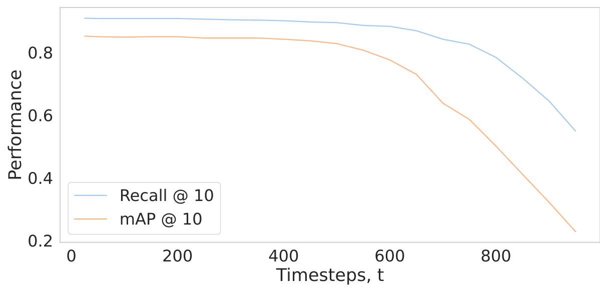


Figure 8. Effects of different timesteps on WikiArt image retrieval.

for certain artists, confirming the observation that style definitions are somewhat fluid and artists sometimes span different styles.

5.3. SHacks Dataset Results

Table 2 compares different approaches on the SHacks dataset. While the recall metric provides less information about the differences, with all of them being 1.00, our precision is substantially better compared to other methods. In particular, IntroStyle outperforms a strong baseline model, CLIP, by 3% in the mAP@5 metric, showing that it is much more effective at retrieving images crafted by "hacked" prompts that indirectly refer to a specific artistic style. This can also be seen by the visualizations of Fig. 6.

5.4. Ablations

In this section, we ablate various parameters and discuss best choices for timestep t, block index idx, and similarity metric. All experiments are performed on WikiArt.

Timestep Fig. 8 shows how style attribution performance, measured by recall and mAP at $k = 10$, varies with the timestep $t \in [25, 950]$ for the implementation of IntroStyle with features of up block index $idx = 1$ and the 2-Wasserstein distance. The performance remains stable until around $t = 400$, dropping significantly after that. We choose $t = 25$ as the default value for the timestep hyperparameter of IntroStyle. Similar observations hold for up block indices $idx \in 0, 2, 3$, which are presented in the supplementary.

Up block index Table 3 compares the performance of IntroStyle with $t = 25$ and the 2-Wasserstein distance for different values of the up block index idx. The best overall performance on both datasets is achieved when $idx = 1$.

Metric	mAP@k 1	10	100	Recall@k 1	10	100
L_2	0.888	0.852	0.521	0.888	0.907	0.938
Gram	0.869	0.835	0.491	0.869	0.885	0.916
JSD	0.888	0.849	0.525	0.888	0.908	0.940
W_2	0.887	0.852	0.525	0.887	0.909	0.941

Table 4. Comparisons on different metrics.

Method	Number of Parameters
DINO ViT-B/16 [3]	21M
CLIP ViT-L [26]	427M
GDA DINO ViT-B [43]	86M
CSD ViT-L [37]	305M
IntroStyle ($idx = 1$)	808M

Table 5. Comparison on the model sizes.

Results also show that IntroStyle is robust to the specific features used to create the feature representation. We also compare the model size for each idx value in supplementary.

Similarity metric Table 4 compares the performance of IntroStyle for the different similarity metrics of Section 3.3, on WikiArt, using $t = 25$ and $idx = 1$. The 2-Wasserstein distance has the best overall performance, albeit it is only marginally better than the Euclidean Distance and the JSD. This shows that IntroStyle has a fairly stable performance with respect to the choice of similarity measure. We choose the 2-Wasserstein distance as the default metric for implementing IntroStyle.

6. Discussion

We presented IntroStyle, a training-free style attribution model that extracts stylistic features and computes similarity scores between images. IntroStyle was shown to outperform existing methods across all metrics. To address the lack of style-specific datasets, we developed SHacks, a synthetic dataset that isolates stylistic elements and enables fine-grained attribution evaluation. IntroStyle's robust performance on both standard and synthetic datasets establishes it as a state-of-the-art method for style attribution, with potential applications in protecting artists' intellectual property rights.

IntroStyle has $2.5\times$ more parameters than CSD (Table 6), which constrains its deployment on memory-limited devices. However, for evaluating generated image style attribution, the model can leverage existing diffusion model features without additional memory overhead. While our current evaluation focuses on Stable Diffusion v2.1, future work could explore compatibility with other pre-trained diffusion models, investigate performance gains through additional linear/MLP layers, and examine applications in diffusion-based style transfer.

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1

[2] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3, 6, 7, 8

[4] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023. 2

[5] Wei-Ta Chu and Yi-Ling Wu. Image style classification based on learnt deep correlation features. *IEEE Transactions on Multimedia*, 20(9):2491–2502, 2018. 3

[6] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8795–8805, 2024. 3

[7] Mathew Dryhurst. AI art and the problem of consent. *ArtReview*, 2023. Accessed: 2024-11-14. 1

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 3, 6, 7

[10] Matt Growcoot. Midjourney founder admits to using a 'hundred million' images without consent. *PetaPixel*, 2022. Accessed: 2024-11-14. 1

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. arxiv e-prints, art. *arXiv preprint arXiv:1911.05722*, 2019. 6, 7

[12] The Sydney Morning Herald. What does the rise of AI mean for the future of art? *The Sydney Morning Herald*, 2022. Accessed: 2024-11-14. 1

[13] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002. 7

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 3, 5

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[17] Jose Antonio Lanz. Greg rutkowski was removed from stable diffusion, but ai artists brought him back. *Decrypt*, 2023. 1

[18] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812, 2021. 3

[19] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3673–3682, 2019. 3

[20] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4990–4998, 2017. 3

[21] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3

[22] OpenAI. Chatgpt, 2024. 5

[23] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 6

[24] Vitali Petsiuk and Kate Saenko. Concept arithmetics for circumventing concept inhibition in diffusion models. In *European Conference on Computer Vision*, pages 309–325. Springer, 2025. 2

[25] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022. 6, 7

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 6, 7, 8

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[28] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative ad-

versarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 3

[29] Reuters News. Ai-created images lose US copyrights in test of new technology. *Reuters*, 2023. Accessed: 2024-11-14. 1

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[31] Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. Aladin: All layer adaptive instance normalization for fine-grained style similarity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11926–11935, 2021. 3, 6

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[33] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arxiv 2015. *arXiv preprint arXiv:1505.00855*, 2015. 1, 6

[34] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1, 3, 5

[35] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023. 2

[36] Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, Bill Freeman, and Mark Matthews. Alchemist: Parametric control of material properties with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24130–24141, 2024. 3

[37] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 2, 3, 4, 6, 7, 8

[38] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[39] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 2, 3, 4

[40] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2024. 2, 3, 4

[41] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023. 2

[42] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended textual conditioning in text-to-image generation. 2023. 2, 3, 4

[43] Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7192–7203, 2023. 2, 3, 4, 6, 7, 8

[44] Sheng-Yu Wang, Aaron Hertzmann, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Data attribution for text-to-image models by unlearning synthesized images. *arXiv preprint arXiv:2406.09408*, 2024. 1, 2, 3

[45] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[46] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15802–15812, 2023. 3, 4

[47] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 3

[48] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 3

[49] Wei Zhang, Chen Cao, Shifeng Chen, Jianzhuang Liu, and Xiaoou Tang. Style transfer via image component analysis. *IEEE Transactions on multimedia*, 15(7):1594–1601, 2013. 2

[50] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023. 3, 4

Supplementary Material

A. Overview

In this supplementary material, we present further details about our methodology and experimental findings. Specifically, we provide an analysis of the hyper-parameter selection for `IntroStyle` features in Sections B and C. We conduct a detailed examination of the t-SNE distribution patterns in Section D. Furthermore, we elaborate on our prompt engineering process utilizing ChatGPT for style isolation in the synthesis of `SHacks` dataset in Section E. Finally, we present additional experimental results and analyses on both the WikiArt and `SHacks` datasets in Section F. Our codes and `SHacks` dataset will be released.

B. Similarity Metrics

In this section, we include the details and formulate the Euclidean, Gram Matrices (using `IntroStyle` features) and Jensen-Shannon Divergence (JSD) metrics as discussed in section 3.3 of the main text.

The L_2 distance (Euclidean distance)

$$L_2(\mu_1, \mu_2)^2 = \|\mu_1 - \mu_2\|_2^2, \tag{5}$$

ignores covariance information and does not have an interpretation as a measure of similarity between probability distributions. This is also the case for the Frobenius norm between the Gram matrices, which is popular in the style transfer literature. It extracts deep features from the two images and then takes the outer product of their corresponding mean vectors μ_1 and μ_2,

$$\text{Gram}(\mu_1, \mu_2) = \|\mu_1\mu_1^T - \mu_2\mu_2^T\|_F. \tag{6}$$

Another popular similarity measure between distributions is the Kullback–Leibler (KL) divergence. For two multivariate Gaussians, it takes the form

$$\text{KL}((\mu_1, \Sigma_1)||(\mu_2, \Sigma_2)) = \tag{7}$$
$$\frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} + \text{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1)\right].$$

Note that the KL divergence is not symmetric. To address this, the Jensen-Shannon Divergence (JSD) averages the KL divergence going in both directions,

$$\text{JSD}(I_1||I_2) = \frac{1}{2}\text{KL}(I_1||I_2) + \frac{1}{2}\text{KL}(I_2||I_1). \tag{8}$$

C. Timestep and Block Index

We study the choices of timestep t and block indices idx of `IntroStyle` by evaluating the image retrieval performance on the WikiArt Dataset. The results are presented in Figs. 9 and 10, showing that best performance is obtained with $t < 500$ and up-block indices $idx = 1$ or $idx = 2$, achieving a balanced trade-off between mAP@10 and recall@10. We chose idx $= 1$ as out default configuration for computational efficiency, as shown in Table 6.

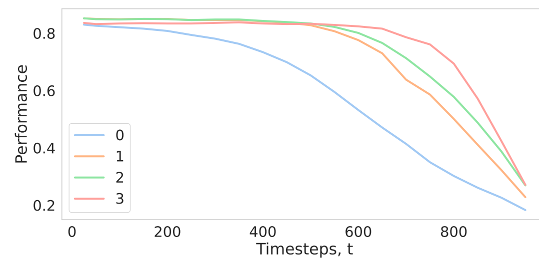


Figure 9. Precision (mAP@10) as a function of timestep t for different up-block indices (idx) on the WikiArt dataset.

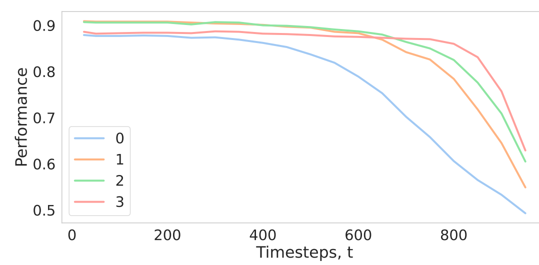


Figure 10. Recall (Recall@10) as a function of timestep t for different up-block indices (idx) on the WikiArt dataset.

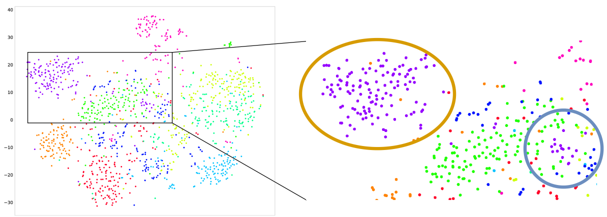
Method	Parameters	Channel (C)
`IntroStyle` (idx = 0)	548M	1280
`IntroStyle` (idx = 1)	808M	1280
`IntroStyle` (idx = 2)	881M	640
`IntroStyle` (idx = 3)	900M	320

Table 6. Comparison of model size needed to compute `IntroStyle` features and channel size for different choices of up-block index (idx).

D. Analysis of t-SNE Distribution

To gain more insight into Figure 7 in the main text, we provide a detailed analysis of the t-SNE Visualization of the mean μ of `IntroStyle` features in Figure 11. In the figure, different styles are indicated by different colors. For example, all purple dots are associated with the same artist.

The purple dots are mainly distributed in two distinct clusters. Visualization of the images from different clusters shows a significant stylistic variation of an artist. These sub-clusters show that an artist often takes different distinctive artistic approaches in different phases of their career. While artworks within a sub-cluster show stronger stylis-



(a) The purple dots are images of *Jacob Jordaens*

(b) Images of *Jacob Jordaens* in the gold sub-cluster.

(c) Images of *Jacob Jordaens* in the blue sub-cluster.

(d) Images of other artists in the blue cluster.

Figure 11. Analysis of t-SNE cluster formation using `IntroStyle` features for a specific artist (*Jacob Jordaens*) whose images are split into two sub-clusters and the images within the blue sub-cluster are stylistically similar to images of other artists in that sub-cluster than those in the gold sub-cluster.

tic similarities, it might differ significantly between sub-clusters. This again emphasizes the discriminative power of `IntroStyle` features, which successfully capture and differentiate between multiple stylistic works from the same artist instead of grouping all works by the artist attribution.

E. Generating Prompts for `SHacks` dataset

As explained in the main text, we curated a comprehensive collection of prompt-image pairs representing 25 influential artists across various artistic movements and periods from the LAION Aesthetics Dataset: `Frida Kahlo, Vincent Van Gogh, Josephine Wall, Gustav Klimt, Takashi Murakami, Pablo Picasso, Claude Monet, Leonid Afremov, Antoine Blanchard, Thomas Kinkade, George Seurat, Katsushika Hokusai, Wassily Kandinsky, Amedeo Modigliani, Alex Gray, Max Ernst, Leonardo Da Vinci, Michelangelo, Gustave Courbet, Isaac Levitan, Andy Warhol, Salvador Dali, Frank Stella, Mary Cassatt, John Singer Sargent.` Using their seminal works as reference queries, we implemented a systematic prompt-generation strategy. For each original prompt, we derived 3 variants of 2 distinct prompt types, Hacked and Innocent, subsequently employing a Stable Diffusion v2.1 to synthesize 12 images per prompt. This methodological approach yielded a richly diverse reference dataset comprising 1,800 images (25 artists $\times$ 6 prompts $\times$ 12 images).

We leveraged the ChatGPT to systematically generate both hacked and innocent prompts.

Hacked Prompt. To create descriptions that effectively mask the original artistic attribution while ensuring consistency and reproducibility in prompt generation, we crafted a base template query structured as follows:

> *"Can you give prompts to obtain a painting of [artist]'s [painting] without mentioning their name or their artwork in the prompt?"*

with `artist` and `painting` being the variable inputs based on the chosen paintings. An example response by ChatGPT, for `Vincent Van Gogh` and his work `Starry Night`, is shown in Figure 12.

Innocent Prompt. The innocent prompts are created by removing style descriptors from the hacked prompt using the following template query.

> *"Simplify the following prompt by removing the style descriptors and keeping the semantic content alone. Prompt : [prompt]"*

with `prompt` referring to the hacked prompt previously generated with an example in shown in Figure 13.

In Figure 14, examples of hacked and innocent prompts for the paintings shown on the left and the resulting synthesized images. Note how the prompts isolate style in various cases.

F. Results on WikiArt and `SHacks` Datasets

Figure 15 and Figure 16 show retrieval results of the proposed `IntroStyle` method on WikiArt. Remarkably, `IntroStyle` is able to identify images of the same style despite a great diversity of content. Note that the images retrieved rarely present images of similar content, indicating the robustness of the proposed approach in focusing on

User

Can you give prompts to obtain a painting of Vincent Van Gogh's "Starry Night" without mentioning their name or their artwork in the prompt?

ChatGPT

To create a painting reminiscent of Vincent van Gogh's "Starry Night" without mentioning his name, you can focus on describing the distinctive elements, style, and atmosphere of the artwork. Here are some prompts that capture its essence:

- "Create a swirling night sky scene with bold, expressive brushstrokes. Include a crescent moon, vibrant stars, and a small village below with a prominent church spire."
- "Paint a dynamic nocturnal landscape featuring a turbulent sky filled with spiraling clouds and radiant stars. Depict a quiet town nestled beneath rolling hills, with cypress trees in the foreground."
- "Design a post-impressionist nightscape with swirling stars and a crescent moon over a small village."

These prompts emphasize the key features of "Starry Night," such as the swirling sky, the bright celestial bodies, the village with its church, and the distinctive brushwork style, without directly referencing the artist.

Figure 12. Hacked prompt generation using ChatGPT.

styles. Furthermore, we show more retrieval results in Figure 17 for the `SHacks` Dataset, where our method works effectively and achieves high retrieval accuracy as well.

The retrieval results, in fact, also demonstrate the effectiveness of the hacked prompts, where images generated by the hacked prompts are indeed similar to the reference image in styles. Our approach successfully retrieves images generated by the hacked prompts, which makes it potentially a strong method for protecting artist styles in real-world applications.

User

Simplify the following prompt by removing the style descriptors and keeping the semantic content alone.
Prompt: Paint a dynamic nocturnal landscape featuring a turbulent sky filled with spiraling clouds and radiant stars. Depict a quiet town nestled beneath rolling hills, with cypress trees in the foreground.

ChatGPT

Here's a simplified version of the prompt, focusing on the semantic content without style descriptors:

- "Paint a night landscape with a sky full of clouds and stars. Show a town below hills, with trees in the front."

Figure 13. Innocent prompt generation using ChatGPT.

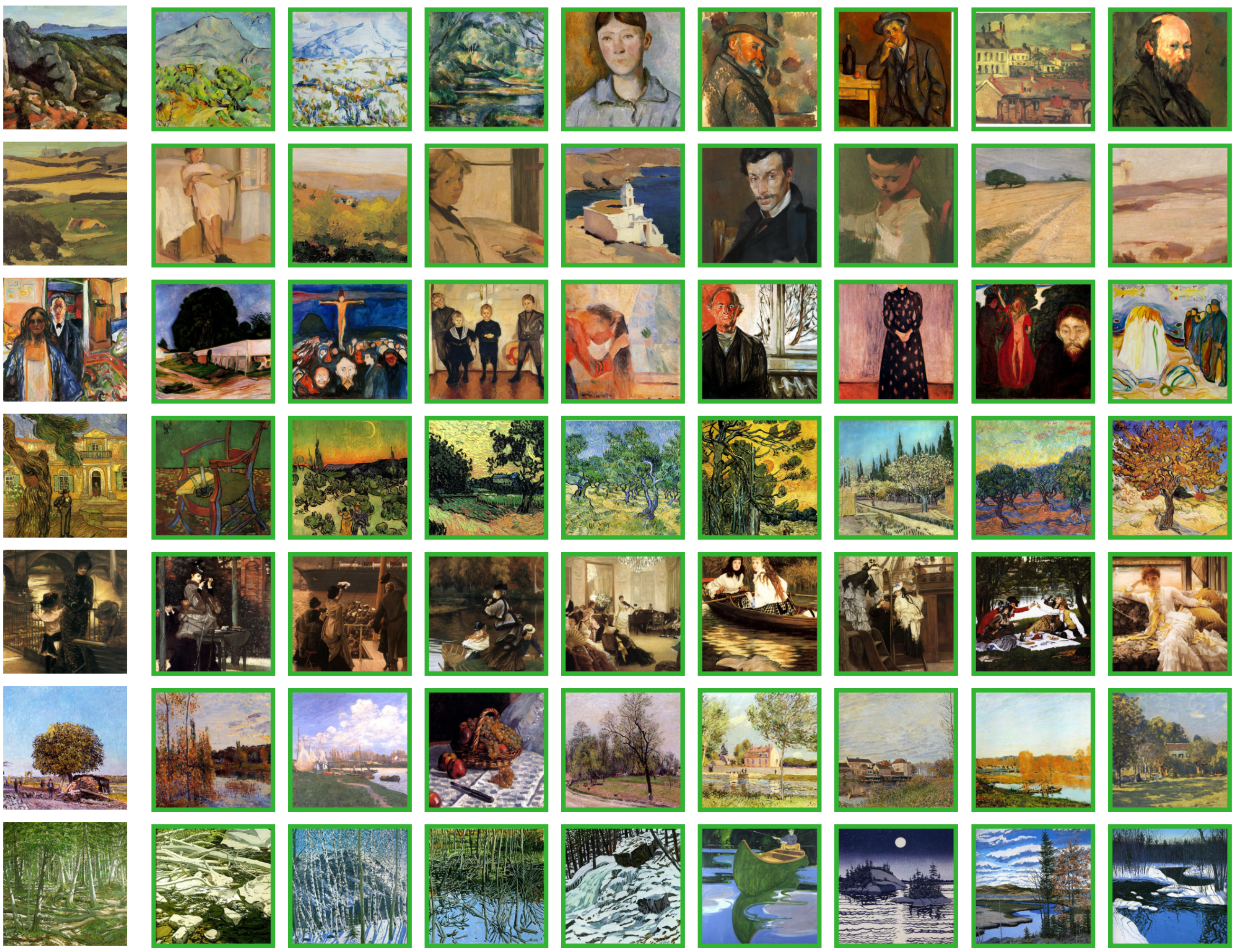
(c.1) Katsushika Hokusai's *The Great Wave off Kanagawa* (c.2) *Design an ukiyo-e style image of a colossal ocean wave with finger-like tendrils of foam. Depict fishermen in traditional boats beneath the wave, and a serene mountain peak on the horizon.* (c.3) *Design an image of a colossal ocean wave with foam.*

(c.1) Vincent Van Gogh's *Starry Night* (c.2) *Design a post-impressionist nightscape with swirling stars and a crescent moon over a small village.* (c.3) *A small village with crescent moon and stars at night sky*

Figure 14. *Style Hacks* (`SHacks`) dataset samples. Each row shows images generated for a reference image by an artist. From left to right: reference image, three images generated by "hacked" prompts, and three images generated by "innocent" prompts.

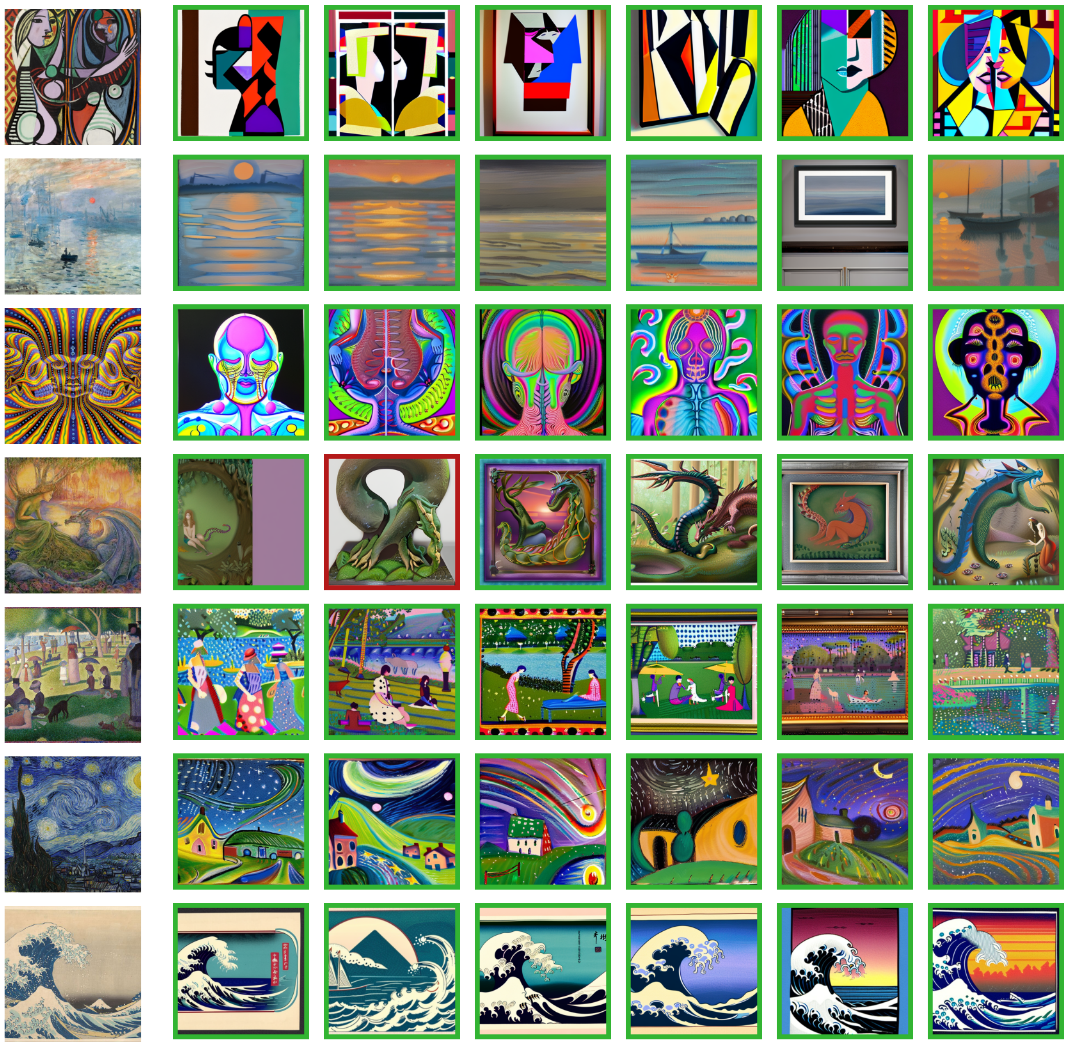
(a) Queries (b) `IntroStyle` (ours)

Figure 15. Additional Image Retrieval on WikiArt Dataset for `IntroStyle` with images ranked highest to lowest from left to right. Green colors indicate correct and red for incorrect retrievals.



(a) Queries (b) `IntroStyle` (ours)

Figure 16. Additional Image Retrieval on WikiArt Dataset for `IntroStyle` with images ranked highest to lowest from left to right. Green colors indicate correct and red for incorrect retrievals.



(a) Queries (b) `IntroStyle` (ours)

Figure 17. Additional Image Retrieval on `SHacks` Dataset for `IntroStyle` with images ranked highest to lowest from left to right. Green colors indicate correct and red for incorrect retrievals.