Discovering and Personalizing Artistic Styles with Generative Models

Matthew Y. Zheng

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science

Pinar Yanardag, Chair
Christopher Thomas
Hoda Eldardiry

May 12, 2025
Blacksburg, Virginia

Keywords: Generative Models, Text-to-Image Generation, Computer Vision, etc.

Discovering and Personalizing Artistic Styles with Generative Models

Matthew Y. Zheng

(ABSTRACT)

Text-to-image models have gained widespread popularity, transforming digital art creation by allowing users to generate highly detailed and imaginative visual content from natural language prompts. These models are now widely adopted across various domains, particularly in the arts, where they enable a broad range of creative expression and make artistic creation more accessible to a wider audience. In this work, we focus on discovering and personalizing emergent artistic styles. By clustering millions of user-generated images from Artbreeder—a platform with over 13 million users—we uncover a rich landscape of previously undocumented unique styles, transcending conventional categories like 'cyberpunk' or 'Picasso,' that reflect the collective creative exploration of users worldwide. Building on these discoveries, we assess personalization methods to align generated content with individual aesthetic preferences and introduce a style recommendation system based on historical behavior. To support this effort, we curate and release `STYLEBREEDER`, a large-scale dataset comprising 6.8 million images and 1.8 million text prompts from 95,000 users, along with clustering annotations and stylistic embeddings. We also introduce the Style Atlas platform, providing public access to 100 curated style LoRA models for user experimentation. Our work demonstrates that AI is not only a tool for generating art, but also a means of discovering and fostering emerging forms of creativity. By openly sharing our data, models, and tools, we aim to support a more diverse, inclusive, and collaborative digital art ecosystem. All resources are available at https://stylebreeder.github.io under a Public Domain (CC0) license.

Discovering and Personalizing Artistic Styles with Generative Models

Matthew Y. Zheng

(GENERAL AUDIENCE ABSTRACT)

Artificial intelligence is transforming the way people create art, making it possible for anyone to generate rich, detailed images simply by describing them with words. With the growing popularity of these tools, millions of artworks are being produced every year, showcasing a vast and diverse range of artistic styles. However, much of this creativity goes beyond traditional categories like "cyberpunk" or "cubism"—users are inventing entirely new styles that reflect a collective, evolving form of digital creativity. In this work, we focus on discovering and personalizing these emerging artistic styles. We study millions of user-generated images from Artbreeder, a platform with over 13 million users worldwide, and group images based on visual similarities to uncover hidden artistic trends that have naturally developed within the community. Beyond simply cataloging these styles, we build tools that recommend styles to users based on their past creations and allow them to create new, personalized artworks that match their own unique tastes. To support broader exploration, we release `STYLEBREEDER`, a large public dataset containing 6.8 million images and 1.8 million text prompts from 95,000 users. We also introduce the Style Atlas platform, where anyone can browse and download curated artistic styles and use them to inspire their own creations. Our work shows that AI can do more than generate art—it can also help uncover and nurture new forms of creativity. By making these resources freely available, we hope to encourage a more diverse, inclusive, and collaborative future for digital art. All data, models, and tools are available at https://stylebreeder.github.io under a Public Domain (CC0) license.

Dedication

To Kirk Lininger, who shaped the person I am today.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Pinar Yanardag, for her invaluable guidance and mentorship throughout my research. To my parents, who have been my greatest supporters. And to my friends at Virginia Tech, who have made this one of the most memorable times of my life.

Contents

List of Figures

List of Tables

List of Abbreviations

DDM Denoising Diffusion Models

GAN Generative Adversarial Networks

LDM Latent Diffusion Models

LoRA Low-Rank Adaptation

PEFT Parameter-Efficient Fine-Tuning

T2I Text-to-Image

Chapter 1

Introduction

Text-to-image models, such as Denoising Diffusion Models (DDMs)[25] and Latent Diffusion Models (LDMs)[49], are becoming increasingly popular and are revolutionizing the landscape of digital art creation. These models have become renowned for their ability to generate high-quality, high-resolution images across a wide range of domains, enabling the production of richly detailed and highly creative visual content [10, 15, 23, 29, 39, 74, 75]. Artists and enthusiasts worldwide are increasingly leveraging these models, using diverse textual prompts to create artworks spanning countless styles, thus democratizing the creative process and making artistic expression more accessible than ever before.

This surge in user-generated content presents an intriguing question: beyond the conventional styles typically prompted by terms like 'cyberpunk' or 'Picasso,' what unique, crowd-sourced styles might exist within such a community? These styles, potentially undocumented and uniquely communal, could offer profound insights into the collective creative psyche of users worldwide. Platforms such as Artbreeder, with its community of over 13 million users and millions of generated images, offer a fertile environment for investigating this question. The immense scale and diversity of user-generated content suggest the existence of unique, crowd-sourced styles that are not formally documented but represent genuine artistic trends born from collaborative exploration.

In this work, we focus on discovering and personalizing these emergent artistic styles. By systematically clustering user-generated images based on stylistic similarity, we uncover a

rich and previously uncharted landscape of aesthetic expressions. These style clusters reveal both individual creative signatures and broader communal trends, offering insights into how new artistic genres evolve in the context of generative AI. Building on these discoveries, we further develop personalization methods that allow users to generate new images aligned with specific styles and showcase a recommendation system that predicts styles users are likely to appreciate based on their historical preferences.

While existing datasets, such as DiffusionDB [69] and TWIGMA [6], have made valuable contributions by cataloging AI-generated content, they often feature a smaller user base, omit the original text prompts, or provide limited stylistic analysis from a visual standpoint. To support our style discovery and personalization efforts, we curate a large-scale dataset from Artbreeder, comprising 6.8 million images, 1.8 million associated text prompts, and contributions from 95,000 unique users. This dataset forms the foundation for our experiments and is released publicly under a CC0 license to encourage future research.

Our contributions are summarized as follows:

- We identify and map emergent, user-generated artistic styles through large-scale clustering of images based on stylistic similarity, offering new insights into community-driven creativity.

- We demonstrate the application of these discovered styles for personalized image generation, benchmarking multiple personalization methods across stylistic fidelity and content relevance.

- We introduce a style recommendation system that personalizes artistic exploration by suggesting styles aligned with individual user preferences, making discovery more targeted and meaningful.

- We present an extensive dataset, `STYLEBREEDER`, from Artbreeder on CC0 license, capturing millions of user-generated images and styles and sharing them with the community to further encourage research in this area.

By shifting focus from isolated prompts to the broader organization of artistic styles, our work highlights the collective creativity of user communities and advances the ability to personalize generative models in a scalable, accessible manner.

Chapter 2

Review of Literature

2.1 Generative Models

Generative models are a class of machine learning techniques aiming to model an underlying data distribution such that they are able to generate new, unseen instances. Differing from discriminative models, which learn to classify or regress a label from a set of features, generative models learn the joint probability of the features and labels i.e. the data distribution itself. Their ability to create realistic data samples using learned representations has elevated generative models at the cutting edge of numerous rapidly advancing fields such as computer vision, natural language processing, and audio processing.

2.1.1 Early Works

The concept of generative modeling predates deep learning. Classical approaches such as mixture models (e.g. Gaussian Mixture Models) [8] and Hidden Markov Models [45] have long been used to model complex probability distributions. However, these earlier methods often lacked the expressiveness to handle data such as paragraphs of text or large-scale images, which are high-dimensional and unstructured in nature.

The emergence of deep neural networks forged the way for generative architectures capable of greater representational power. Early breakthroughs such as Restricted Boltzmann Machines

[21] and Deep Belief Networks [21] showcased the potential that deep learning in generative tasks possessed, but were later eclipsed by more scalable architectures.

2.1.2 Modern Generative Models

Following these early advancements, several major paradigms for deep generative models have achieved widespread attention:

1. Variational Autoencoders (VAEs): Proposed by Kingma and Welling in 2013, VAEs [30] presented a powerful framework leveraging latent variables to learn continuous representations of high-dimensional input. VAEs approximate the posterior distribution of latents with a parametric encoder and use a decoder to reconstruct samples. The stochastic "bottleneck" enables VAEs to reflect variations in data while remaining computationally tractable. VAEs excel in learning interpetable and structured latent spaces, enabling tasks such as conditional generation and interpolation between data points.

2. Generative Adversarial Networks (GANs): Introduced by Goodfellow et al. in 2014 [16], GANs revolutionized generative modeling by proposing adversarial training. Consisting of 2 opposing models, a generator that produces fake samples and a discriminator that distinguishes real samples from fake ones, GANs use minimax optimization objective to learn to produce indistinguishable fake samples. They have shown great success in generating realistic images, such as faces, however, they can be difficult to train due to mode collapse and instability stemming from the adversarial objective.

3. Diffusion Models: Diffusion-based approaches [24, 59] have emerged as a highly effective method for generation. Through a forward diffusion process, they gradually add

noise to data, and then learn a reverse denoising process to invert the noise. Diffusion models have demonstrated strong performance in generating both high-quality and diverse images along with greater stability in their training dynamics compared to the pitfalls of GANs.

2.1.3 Parameter-Efficient Fine Tuning

A defining characteristic of modern generative modeling is the progressive emphasis on augmenting scale in data and network parameters. With models such as Stable Diffusion and GPT-4 employing hundreds of millions to billions of parameters [5, 44], this increase in scale has empowered more expressive representations and improved sample quality, but also imposes substantial demands in computational resources.

As model sizes continue to grow, full fine-tuning—where all model parameters are updated for each new task—becomes increasingly impractical. This challenge has motivated the development of Parameter-Efficient Fine-Tuning (PEFT) methods [26, 31, 33**?**], which aim to adapt large pre-trained models to new tasks by modifying only a small subset of parameters while keeping the majority of the model frozen. PEFT techniques significantly reduce the computational overhead typically associated with model customization, making personalization and downstream adaptation more accessible.

One widely adopted PEFT approach is Low-Rank Adaptation (LoRA) [26], which injects trainable low-rank matrices into the attention layers of the model. At a high level, LoRA modifies the weight update process in neural networks by introducing a low-rank decomposition. Instead of updating the full weight matrix $W \in \mathbb{R}^{d\times k}$ during fine-tuning, LoRA freezes W and learns two smaller matrices, $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$, where $r \ll \min(d, k)$. The adapted weight during training is expressed as:

$$W' = W + \Delta W, \quad \Delta W = AB$$

This formulation allows the model to capture task-specific changes through ΔW while keeping the original pre-trained weights intact. By constraining ΔW to be low-rank, LoRA significantly reduces the number of trainable parameters and the memory footprint of fine-tuning, while still allowing sufficient expressiveness for adaptation. LoRA has been particularly effective for adapting large language models and diffusion models without sacrificing generation quality [26, 55, 70].

2.2 Diffusion Models for Image Generation

2.2.1 Overview

Diffusion models have recently emerged as a powerful class of generative modeling frameworks capable of generating high-fidelity and diverse image outputs. Fundamentally, these models corrupt training data over a series of time steps by adding noise (forward process) and learning a neural network to remove that noise step-by-step (reverse process). First introduced by Ho et al. in 2020 as Denoising Diffusion Probabilistic Models (DDPM) [24], diffusion models have since improved upon, and in many settings surpassed, the performance of GANs in generated image quality and training stability.

2.2.2 The Forward-Backward Process

The foundation of diffusion models consists of two processes: a forward (noising) process and a reverse (denoising) process:

Forward Process

During the forward diffusion process, an image from the training dataset $\mathbf{x_0}$ is progressively adding noise from a Gaussian distribution over T time steps. This forward noising process is defined as a Markov chain:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}),$$

where each transition is given by:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}),$$

where β_t denotes a noise variance schedule. After T time steps, the forward process destroys the image sample until it is pure noise.

Reverse Process

The goal of the reverse process is to learn how to invert this corruption by progressively denoising $\mathbf{x}_T$ back to a clean sample $\mathbf{x}_0$. Formally, the reverse process is modeled as another Markov chain:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t),$$

where $p(\mathbf{x}_T)$ is typically a standard Gaussian prior, and $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is parameterized by a neural network with parameters θ.

Each reverse transition is modeled as a Gaussian distribution:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)),$$

where $\mu_\theta(\mathbf{x}_t, t)$ and $\Sigma_\theta(\mathbf{x}_t, t)$ are the predicted mean and variance at each timestep. In practice, the variance Σ_θ is often fixed or learned separately, and the model focuses on predicting the mean.

Training is performed by minimizing a variational bound on the negative log-likelihood. In DDPM [25], the model is simplified to predict the noise ϵ added during the forward process, rather than directly predicting the mean μ_θ. Given a sample $\mathbf{x}_t$, the model predicts $\epsilon_\theta(\mathbf{x}_t, t)$, and the mean can be reconstructed as:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right),$$

where α_t and $\bar{\alpha}_t$ are functions of the noise schedule.

At inference time, the model starts with a random Gaussian sample $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively applies the learned denoising transitions $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ for $t = T, T-1, ..., 1$. Through this stepwise refinement, the noisy input is progressively denoised until a coherent, high-fidelity image is produced.

2.2.3 Variations and Advancements

Following the initial formulation of DDPMs, several refinements have been introduced to enhance the efficiency, stability, and sample quality of diffusion models. Key developments include:

1. **Improved Noise Scheduling:** The forward noise schedule, determined by the vari-

ance parameters β_t, significantly influences training dynamics and sampling quality. In the original DDPM formulation, a linear schedule was used. Subsequent work has shown that alternative schedules, such as cosine or learned noise schedules [43], can accelerate convergence, improve sample fidelity, and stabilize training. These schedules control how quickly signal-to-noise decays across time steps, which directly impacts the model's ability to reconstruct fine details during the denoising process.

2. **Denoising Diffusion Implicit Models (DDIM):** DDIM [60] introduced a deterministic alternative to the stochastic sampling procedure of DDPMs. Instead of sampling $\mathbf{x}_{t-1}$ with added noise at each step, DDIM formulates the reverse process as a deterministic mapping conditioned on the predicted noise $\epsilon_\theta(\mathbf{x}_t, t)$. Specifically, the update rule is given by:

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sigma_t \mathbf{z},$$

where σ_t controls the amount of stochasticity and $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. In the deterministic case where $\sigma_t = 0$, the trajectory becomes fully non-stochastic. DDIM preserves the underlying structure of the denoising trajectory while allowing samples to be generated with significantly fewer steps than required by traditional DDPMs, resulting in faster sampling without substantial degradation in image quality.

3. **Latent Diffusion Models (LDM):** Latent Diffusion Models [49] propose an architectural refinement that improves efficiency by applying the diffusion process in a lower-dimensional latent space rather than directly in pixel space. In this approach, a Variational Autoencoder (VAE) is first trained to encode images into a latent space $\mathcal{Z}$, where the diffusion model learns to operate. Formally, an image $\mathbf{x}_0$ is mapped to a latent vector $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ using an encoder $\mathcal{E}$, and the denoising process is per-

formed on $\mathbf{z}_t$ within $\mathcal{Z}$. After sampling, the decoder $\mathcal{D}$ reconstructs the final image as $\hat{\mathbf{x}}_0 = \mathcal{D}(\mathbf{z}_0)$. Operating in the latent space significantly reduces computational cost and memory footprint while preserving high fidelity in the final generated images. Latent diffusion enables scalable high-resolution image synthesis that would otherwise be infeasible with pixel-space diffusion models.

2.3 Text-to-Image Generation

Text-to-image (T2I) generation [44, 46, 49, 52] seeks to incorporate human language descriptions into semantically and visually aligned images. At a high level, the process is composed into two core components: (1) a text encoder that transforms text prompts into an embedding space, and (2) a generative network that utilizes those embeddings to condition the generation process.

Contemporary approaches to text encoding typically rely on large language models (LLMs) or multimodal embeddings derived using a transformer to convert text into numerical token representations. For example, a text encoder (like CLIP [41]) produces an embedding capturing semantic and syntactic information in the prompt, which is passed to the generative model.

Early GAN-based systems incorporated text embeddings into the generator and discriminator [48], laying the groundwork for T2I generation. Although these versions revealed conditioning on text for image synthesis was feasible, they struggled to capture complex semantics and scenarios. Progressively, researchers introduced more refined architectures (e.g. StackGAN [73], AttnGAN [71]) that improved image details and relied on attention mechanisms to focus on prompt segments. Recent T2I frameworks, including Stable Diffusion and FLUX [12, 44], have adopted diffusion-based generative models, where at each denoising

step, the model is conditioned on the text embedding to guide the progressive refinement of noise into a coherent image. Diffusion-based methods have significantly advanced the quality, versatility, and controllability of T2I generation. A key innovation enabling fine-grained conditioning is the integration of text embeddings through cross-attention mechanisms within the generative model.

Formally, cross-attention integrates the text features into the model by modifying the intermediate feature maps at each denoising step. Given a latent feature map $\mathbf{h} \in \mathbb{R}^{n \times d}$ from the diffusion model (where n is the number of spatial tokens and d is the feature dimension) and a text embedding sequence $\mathbf{e} \in \mathbb{R}^{m \times d}$ (with m text tokens), the cross-attention mechanism first projects them into query, key, and value representations:

$$Q = \mathbf{h}W_Q, \quad K = \mathbf{e}W_K, \quad V = \mathbf{e}W_V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are learned projection matrices.

The cross-attention output is then computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V.$$

Here, the latent queries attend to the text keys and values, allowing the diffusion model to adapt its visual features based on the semantic content of the prompt at each denoising step.

This conditioning via cross-attention allows the model to flexibly align different regions of the generated image with specific elements of the text description, producing outputs that are both high-resolution and semantically faithful. By injecting language-derived constraints directly into the denoising process, diffusion-based T2I models [46, 49, 52] achieve state-of-the-art performance in generating high-quality and visually aligned images.

As T2I models continue to improve, they have become indispensable tools for creative tasks involving image synthesis, manipulation, and editing. Prominent diffusion-based text-to-image models [2, 24, 60] are increasingly used for guided image synthesis [10, 23, 75] and complex image editing applications [3, 74].

2.4 Artwork Datasets

Traditional artwork datasets (see Supplemental Materials for a comparison) have primarily focused on artwork classification and attribute prediction. However, these datasets often exhibit limitations, like skewed class distributions and unsuitable classes for image synthesis, when employed for artwork synthesis. To address these shortcomings in evaluating artwork synthesis, specialized subsets have been curated to better suit the task. For instance, [64] and [76] derived datasets by scraping WikiArt images. Despite these efforts, such datasets still face many challenges related to variable image quality, imbalanced distributions, and others. ArtBench-10 [35] attempted to rectify these issues by introducing a class-balanced and cleanly annotated benchmark. However, it only provides ten classes.

Recent advancements in text-to-image synthesis have spurred the development of AI-generated datasets like DiffusionDB [69], and Midjourney Kaggle [40], which contain millions of image-text pairs generated by models such as Stable Diffusion and Midjourney. These datasets, while groundbreaking, tend to be limited in stylistic diversity and are skewed towards specific user groups, reflecting data collected from constrained environments and short time frames. The main distinction between our dataset and those datasets lies in the duration over which the images, along with the magnitude of the images. While the DiffusionDB dataset covers a brief period of just 12 days in August 2022, our dataset extends across a much longer time frame, spanning 18 months from July 2022 to May 2024. This extensive duration provides a

significant advantage for in-depth studies into the evolution and dynamics of visual trends, artistic styles, and thematic content. Researchers have tailored other datasets to investigate certain themes [4, 32, 38, 57, 61, 68, 72], but these are domain-specific and lack breadth. TWIGMA [6] captured multiple years of generated images scraped from X (formerly Twitter) but does not include the prompts that were used to generate these images, instead relying on inferred BLIP [34] captions. As shown in Tab. 3.1, our dataset not only includes the original prompts used to produce images but also contains images generated by multiple models while encapsulating a long time frame.

2.5 Personalized Generation in Diffusion Models

Personalization techniques have been proposed to enable pre-trained text-to-image models to generate novel concepts based on a small set of images. DreamBooth [50] fine-tunes the full T2I model, which yields more detailed and expressive outputs. However, due to the large scale of these models, full fine-tuning is an expensive task that requires substantial amounts of memory. Different methods attempt to work around this challenge for both style and content representations. Custom Diffusion [31] attempts multi-concept learning but requires expensive joint training and struggles with style disentanglement. Textual Inversion [14] learns a new token embedding to represent a subject or style without altering the original model parameters, while StyleDrop [56] utilizes adapter tuning to only train a subset of weights for style adaptation. Low-Rank Adaptation (LoRA) [26] is a Parameter Efficient Fine-Tuning (PEFT) technique frequently used to fine-tune T2I models to generate images of a desired style. EDLoRA [18] proposes a layerwise embedding and multi-word representation when training a LoRA model.

Chapter 3

Stylebreeder Dataset

We collect STYLEBREEDER by scraping images from the Artbreeder website. We choose Artbreeder since it is one of the most popular platforms for art generation, supporting various text-to-image models. Artbreeder enables users to create images using text prompts, offering controls over various settings, such as the strength (guidance scale) of the text's influence on the generated image, seed values, model type, and other hyperparameters. Since its rise in popularity within the artistic community in 2018, Artbreeder has become known for its bias towards generating artistic images. This predisposition towards artistic styles is a primary reason we concentrated our focus on this area. Additionally, all images on Artbreeder are covered by a CC0 license[1], which allows for unrestricted use for any purpose [2]. We collect metadata along with the images, which include text prompts (positive and negative), usernames, and hyperparameters. We provide additional features such as NSFW scores for each image and text prompts (see Fig. 3.1).

3.1 Comparisons with Other Datasets

Table 3.2 features several prominent image synthesis benchmarks. Certain benchmarks focus on single-category image datasets for unconditional image synthesis evaluation, such as face datasets like CelebA [36] and FFHQ [27] (human faces) and MetFaces [28] (face

[1] https://creativecommons.org/public-domain/cc0
[2] https://www.artbreeder.com/terms.pdf

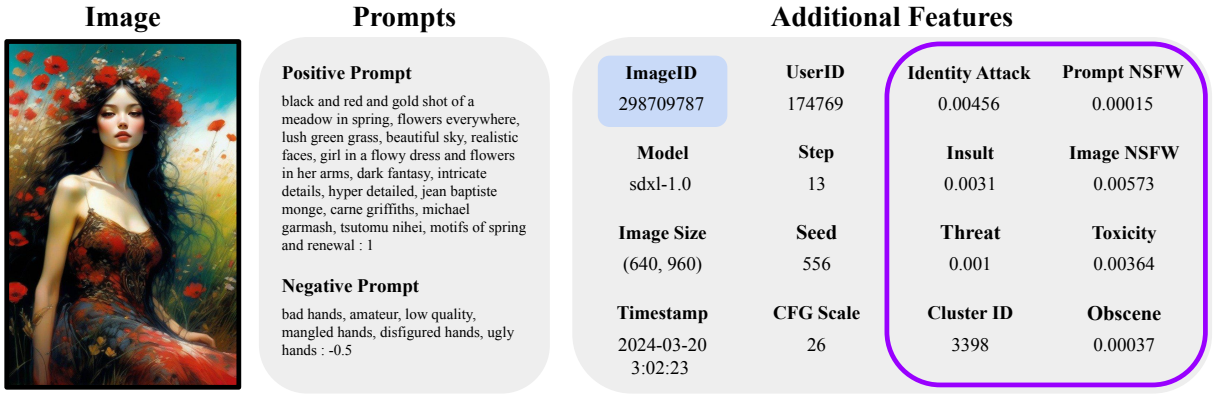
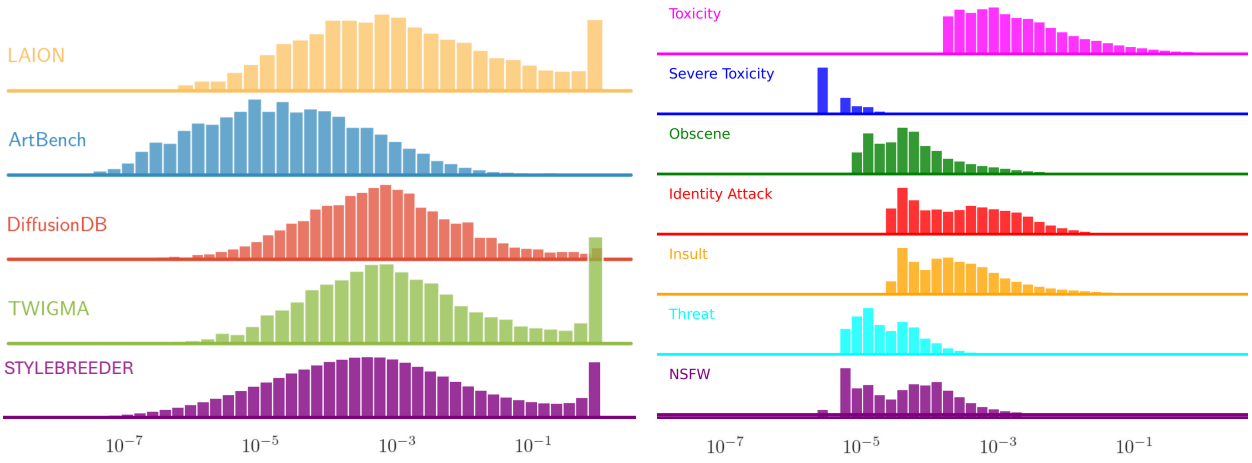


Figure 3.1: Our dataset comprises 6.8M images generated by 95,000 unique users, accompanied by 1.8M text prompts from July 2022 to May 2024. It includes detailed metadata such as Positive Prompt, Negative Prompt, UserID, Timestamp, and Image Size. Additionally, we supply model-related hyperparameters, including Model Type, Seed, Step, and CFG Scale. Note that the disparity in prompts and images arises because different images can be generated from the same text prompt when varying hyperparameters. We also offer further metadata like Cluster ID, along with scores for Prompt NSFW, Image NSFW, and Toxicity computed using state-of-the-art models [19, 20].



(a) NSFW scores for images

(b) NSFW and other scores for text prompts

Figure 3.2: (a) Predicted NSFW scores across LAION [54], Artbench [35], DiffusionDB [69] and TWIGMA [6], STYLEBREEDER (Ours) on images[3], computed with [20] (higher score indicates more NSFW content). (b) Predicted NSFW, Toxicity, Severe Toxicity, Identity Attack, Insult, and Threat scores across on text prompts, computed with [19] on STYLEBREEDER.

Table 3.1: A comparison of our dataset to other AI-generated image datasets

Name	Source	# Images	Year	Original Prompt Included	Multiple Models
DiffusionDB	SD Discord	14,000,000	Aug 2022	✓	✗
Midjourney Kaggle	Midjourney	250,000	Jun 2022-Jul 2022	✓	✗
TWIGMA	Twitter	800,000	Jan 2020-Mar 2023	✗	✓
STYLEBREEDER (Ours)	Artbreeder	6,818,217	Jul 2022-May 2024	✓	✓

artworks). There are also multi-class datasets like STL-10 [7] and ImageNet [9], but these datasets predominantly consist of photographic images with a limited artwork representation. ArtBench-10 [35] is primarily composed of artwork. However, with only ten classes, it heavily skews toward artwork of North American, European, and East Asian origin and fails to encompass digital and modern art. TWIGMA [6] presents an AI-generated art dataset but suffers from a lack of original generation prompts and a significantly smaller volume of images. The main distinction between our dataset and other AI-generated datasets of similar scale, such as DiffusionDB [69], lies in the duration over which images were generated. While DiffusionDB covers images generated during a 12-day period in August 2022, STYLEBREEDER extends across a much longer time frame, spanning 18 months from July 2022 to May 2024. This extensive duration provides a significant advantage for in-depth studies into the evolution and dynamics of visual trends, artistic styles, and thematic content. By covering a broader range of temporal variations, our dataset allows for a more detailed analysis of how generative models respond to changing cultural or seasonal influences over time.

3.2 User Statistics

Our dataset comprises 95,479 unique users, each generating an average of 72.41 images. Figure 3.3 (a) illustrates the distribution of images per user, showing that the majority of users produced fewer than 1,000 images, although a few power users created a significantly larger number of images. All user IDs in our dataset have been anonymized to ensure users'

Table 3.2: Summary of Datasets

Name	Min Resolution	Max Resolution	# Images	Domain
MetFaces [28]	(1024, 1024)	(1024, 1024)	1,336	Faces (art)
STL-10 [7]	(96, 96)	(96, 96)	13,000	Objects
ArtBench-10 [35]	(32, 32)	(10629, 7437)	60,000	Artworks
FFHQ [27]	(256, 256)	(1024, 1024)	70,000	Faces (Flickr)
CelebA [36]	(64, 64)	(1024, 1024)	202,599	Faces (celebrities)
TWIGMA [6]	512^2	Varied	800,000	AI artworks
ImageNet [9]	(32, 32)	(256, 256)	1,431,167	Objects
DiffusionDB [69]	(512, 512)	Varied	14,000,000	AI artworks
STYLEBREEDER (Ours)	(512, 512)	(1280, 986)	6,818,217	AI artworks

privacy.

3.3 Model Statistics

Our dataset represents a wide variety of text-to-image diffusion models, including Stable Diffusion 1.5 (74.9%), SD-XL 1.0 (13.1%), Stable Diffusion 1.3 (8.8%), Stable Diffusion 1.4 (1.3%), Stable Diffusion 1.5-free (1.1%) and ControlNet 1.5 (0.8%). These models differ in their capabilities and the quality of their generated outputs, with recent models often supporting higher resolutions that deliver finer details and more complex visuals. This variety in models used to generate images provides a valuable opportunity to explore their differences, such as variations in artistic expression, the nuances in image quality, and potential biases inherent in each model. Stable Diffusion 1.5 is the most frequently used model in our dataset, followed by SD-XL, which has gained popularity due to its ability to generate high-quality images. The resolution of the images ranges from 512×512 to 1280×896 based on the model employed. Additionally, our dataset provides details on key hyperparameters like seed, CFG guidance scale—which dictates the extent to which the image generation process adheres to the text prompt—and step size in the diffusion model. Users are able to generate varying

images using identical text prompts by adjusting these parameters. How different configurations affect the resulting images introduces deeper insights into the model's behavior and its sensitivity to these parameters. This enables a deeper understanding of how subtle changes in input or settings can significantly alter the characteristics of generated images, providing valuable perspectives on the underlying generative processes.

3.4 Text Prompts

We collect both positive and negative text prompts for each image in our dataset. The average prompt length is 60 words, as shown in Fig. 3.3 (b). Common words in positive prompts, such as 'painting', 'realistic', and 'digital', reveal semantic information about the desired styles of images (see Fig. 3.3 (c)) while common words in negative prompts like 'ugly' and 'deformed' indicate the undesired features of the generated images (see Fig. 3.3 (d)). Furthermore, we observe that users often incorporate artist names in their text prompts to specify desired styles, a common practice employed by the generative art community. To quantify this trend, we analyze using BERT NER [65] to identify unique artist names in the dataset. Our findings highlight a significant occurrence of artist names, with top mentions including 'Ilya Kuvshinov', a Russian illustrator, featured in 208K text prompts, and 'Akihiko Yoshida', a Japanese video game artist, appearing in 81K prompts. Given that these artists may not permit the use of their artistic styles, we offer a form on our website allowing artists to opt-out, ensuring their styles are not replicated without their consent, as outlined in Section 5.1.

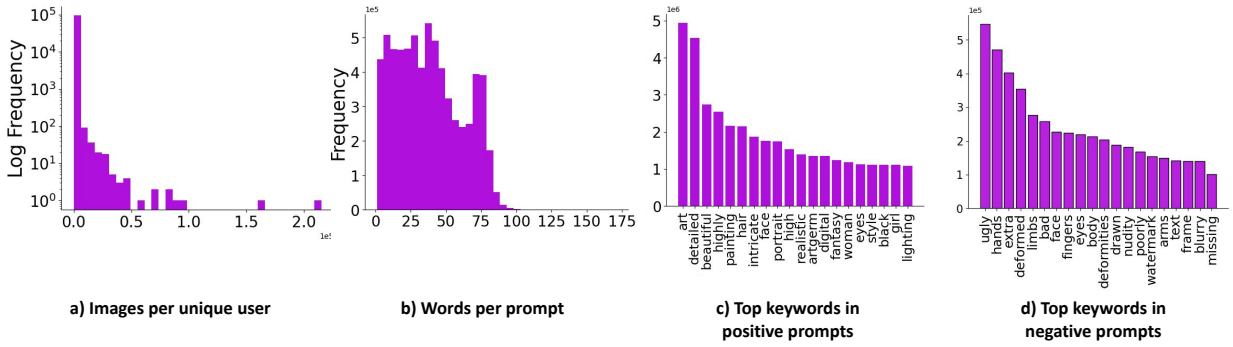


Figure 3.3: Most unique users have fewer than 1000 images generated. The average number of words in a prompt is less than 60 words. Common keywords for positive prompts include 'painting', 'realistic', and 'digital' reveal semantic information about the style of desired images. Common keywords in negative prompts, such as 'ugly' and 'deformed,' indicate undesired features of generated images.

Chapter 4

Discovering and Personalizing Artistic Styles

In this section, we present a comprehensive study AI generated content through a set of carefully designed experiments. We define three key tasks using our newly curated dataset: (1) identifying and categorizing diverse artistic styles through clustering based on stylistic similarities among images, (2) generating personalized images that are tailored to align with an individual's preferred artistic styles, and (3) recommending styles to users based on their historical generation patterns. Together, these tasks illustrate the rich potential of Stylebreeder for advancing research and applications in personalized generative art. We further conduct an exploration of our dataset to reveal detailed insights into the landscape of AI-generated content.

4.1 Experimental Setup

To ensure a fair and robust evaluation across multiple methods, we utilize both the official implementations as well as the HuggingFace Diffusers [67] library for training and inference. Specifically, we experiment with four prominent personalization methods: Textual Inversion [14], LoRA with DreamBooth [26, 50], Custom Diffusion [31], and EDLoRA [18].

For each method, we maintain consistent experimental conditions by using the same set of

random seeds and closely following the hyperparameter settings recommended by the authors in their original publications. For example, Textual Inversion is trained at a resolution of 512×512 pixels for a total of 3000 optimization steps with a learning rate of 5×10^{-4}. In contrast, LoRA with DreamBooth and EDLoRA are both trained at the same resolution for 800 steps, utilizing a slightly smaller learning rate of 1×10^{-4}; in addition, LoRA is configured with a rank of 32 to control the parameter bottleneck during adaptation. Custom Diffusion, which employs a more conservative fine-tuning strategy, is trained for 250 steps with a learning rate of 1×10^{-5}.

To evaluate the quality of the generated images, we compute DINO and CLIP-based scores, as summarized in Tab. 4.2. These metrics are computed over 20,000 generated images for each method to ensure statistical significance, with the standard deviations provided for transparency. All experiments are conducted on 8 NVIDIA L40 GPUs, providing sufficient computational resources for large-scale evaluations.

For style recommendation experiments, we utilize a larger corpus of 96,000 images generated by 1,434 users. We split the dataset into 80% training and 20% test partitions, and employ a matrix factorization approach using the Surprise [62] library. We optimize models using a learning rate of 5×10^{-3} and apply a regularization term of 2×10^{-2} to mitigate overfitting. Additionally, we conduct 5-fold cross-validation to further assess the generalizability of the recommendation models.

4.2 Discovering Diverse Artistic Styles

The rising popularity of generated art has led to a vast array of user-generated content showcasing a diverse spectrum of artistic styles. However, a key challenge remains: how can we effectively discover and categorize these styles? We aim to cluster user-generated images

into stylistically similar groups to uncover unique styles. Formally, our dataset, denoted as $\mathcal{D}$, consists of N images. We employ a pre-trained text-to-image model M_Θ, parameterized by weights θ and a token embedding space V. Firstly, we convert the images into a set of N style embeddings using a state-of-the-art feature extractor F, specifically CSD [58], which uses a Vision Transformer (ViT) [11] backbone to map images into a d-dimensional vector space representing their style descriptors. CSD has shown superior performance in style-matching tasks across datasets like DomainNet, WikiArt, and LAION-Styles, outperforming models like DINO, which focus more on semantic content. This conversion results in a set of embeddings $Z = \cup_N F(\mathbf{x}_i)$, where each image, $\mathbf{x}_i$, is embedded in a high-dimensional semantic embedding space. These embeddings are then clustered into $k = 10000$ groups using the K-Means++ [1] algorithm, which utilizes cosine similarity to ensure cluster cohesion. To determine the optimal number of clusters, we employ the silhouette score method across various cluster sizes: 50, 100, 500, 1000, 2000, 5000, 10000, 20000.

We leverage the silhouette score as our primary criterion to determine the optimal number of clusters, selecting $k = 10000$ as the best configuration. As shown in Tab. 4.1, we also experimented with 20000 clusters; however, increasing the number of clusters did not result in a significant deviation or improvement in the silhouette score compared to the 10000-cluster configuration. This outcome suggests that 10000 clusters provide a balance that yields the most meaningful, distinct, and interpretable categorization of the diverse styles present in our dataset, without introducing unnecessary fragmentation or overfitting.

Our primary objective with this clustering is to effectively capture and group artistic styles within each cluster using the CSD feature extractor [58]. We hypothesize that a substantial fraction of the resulting clusters correspond to either individual artists with uniquely identifiable styles or small groups of artists whose styles exhibit significant visual or thematic similarity. To investigate this hypothesis more rigorously, we conduct a detailed analysis

Table 4.1: Number of clusters k and Silhouette Score

# Clusters	Silhouette Score
50	0.032
100	0.043
500	0.054
1000	0.064
2000	0.078
5000	0.087
10000	0.110
20000	0.111

examining the number of unique artist names assigned within clusters, revealing notable patterns in the distribution and dominance of artistic representation.

Specifically, we analyze the dominance of individual artists or small groups of artists within clusters by inspecting the proportion of images contributed by the top artists in each cluster. Our observations can be summarized as follows:

- 1551 clusters are dominated by a single artist: This means that in these clusters, over 50% of the data points belong to a single artist, highlighting a strong association between the cluster and that artist's distinct style.
- 2345 clusters are dominated by two artists: This suggests that these clusters capture stylistic similarities between two artists, potentially representing shared influences, overlapping techniques, or a broader stylistic movement encompassing both artists.
- 1467 clusters are dominated by the three artists: This further expands the scope of shared stylistic traits, indicating potential sub-genres or broader artistic trends encompassing a small group of artists.
- 884 clusters are dominated by four artists: This reinforces the trend of clusters capturing shared stylistic qualities among a small group of artists, suggesting the presence of

broader artistic movements or schools of thought.

These results reveal a fascinating dynamic between highly individualized artistic expression and broader stylistic trends. While a substantial portion of clusters (1551 out of 10000, or approximately 15.5%) strongly represent the distinct style of a single artist, a larger overall fraction — almost 52% when including clusters dominated by two, three, or four artists — reflects the presence of shared visual motifs, techniques, and aesthetics among multiple creators. This interplay between individual distinctiveness and collective stylistic trends highlights the complexity and richness of the artistic ecosystem captured in our dataset.

Moreover, this experiment further validates that the choice of 10000 clusters indeed maximizes the silhouette score, providing strong evidence of optimal internal similarity within clusters and meaningful external dissimilarity across different clusters. To visually illustrate the structure of the learned style space, we present in Figure 4.4(c) a 2D projection of the CSD embedding space generated via t-SNE [22]. The resulting visualization reveals distinct patterns: some clusters are closely grouped, reflecting subtle stylistic similarities, while others are clearly isolated, indicating significant stylistic variation and confirming the effectiveness of our clustering strategy in organizing the stylistic diversity of the dataset.

4.2.1 Personalized Image Generation Based on Style

Personalized image generation is crucial to creative AI applications, enabling users to produce content that is uniquely aligned with their personal aesthetic preferences. In this task, we benchmark four leading personalization techniques—Textual Inversion [14], LoRA with DreamBooth [26, 50], Custom Diffusion [31], and EDLoRA [18]—on their ability to faithfully adapt to and generate images in specific artistic styles discovered from our clustering.

For a thorough evaluation, we randomly select 40 clusters from our previously discovered set.

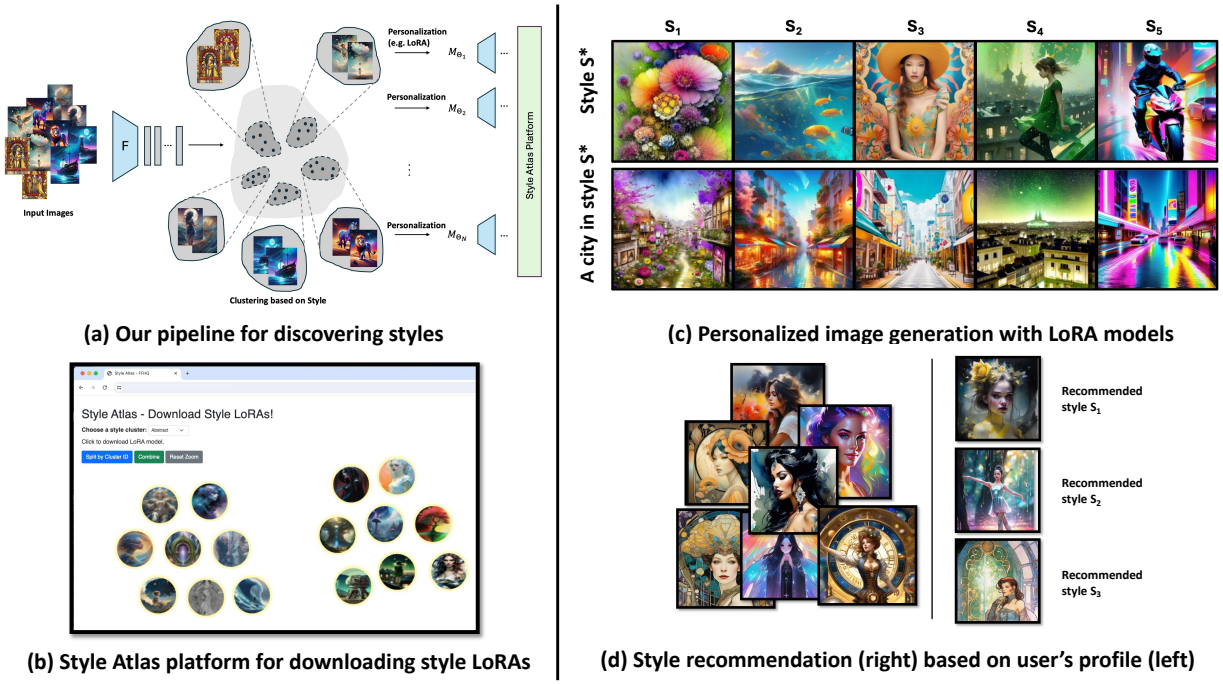


Figure 4.1: (a) An illustration of our pipeline: we cluster input images by stylistic similarity and employ a personalization method, such as LoRA, to train personalized models aligned with specific styles. (b) Users can download style LoRA models from the Style Atlas platform. (c) Users can generate personalized images using LoRA models where `Style S*` represents an example image from the cluster. (d) We recommend top styles to users based on the images they have previously generated. This personalized approach helps tailor style suggestions to each user's unique preferences.

For each cluster, we generate 50 images for each of ten different text prompts, resulting in a dataset of 500 generated images per cluster, per method. This experimental design ensures a diverse set of outputs for each style-method pairing and enables rigorous quantitative analysis.

We evaluate the quality of personalization using three complementary metrics. CLIP-T measures the text-image alignment between the generated image and its associated text prompt, reflecting how well the model preserves textual semantics. CLIP-I measures the similarity between the generated image and the style cluster centroid, quantifying how well the output aligns stylistically with the intended cluster. Finally, DINO captures semantic consistency

Table 4.2: Benchmark results for state-of-the-art personalized image generator models.

		Textual Inversion	LoRA w/DreamBooth	EDLoRA	Custom-Diffusion
CLIP-I	Avg.	0.6869 ± 0.10	0.6299 ± 0.11	**0.6957 ± 0.11**	0.5917 ± 0.12
	Min.	0.6166 ± 0.10	0.5654 ± 0.11	**0.6214 ± 0.11**	0.5324 ± 0.11
	Max.	0.7428 ± 0.10	0.6831 ± 0.11	**0.7521 ± 0.12**	0.6440 ± 0.12
CLIP-T	Avg.	0.1857 ± 0.02	**0.1896 ± 0.02**	0.1822 ± 0.01	0.1809 ± 0.02
	Min.	0.1555 ± 0.02	**0.1573 ± 0.02**	0.1527 ± 0.01	0.1486 ± 0.02
	Max.	0.2392 ± 0.03	**0.2663 ± 0.03**	0.2389 ± 0.03	0.2585 ± 0.03
DINO	Avg.	0.3801 ± 0.15	0.2668 ± 0.17	**0.4125 ± 0.18**	0.2546 ± 0.17
	Min.	0.2581 ± 0.13	0.1682 ± 0.14	**0.2790 ± 0.15**	0.1634 ± 0.14
	Max.	0.4838 ± 0.17	0.3585 ± 0.19	**0.5246 ± 0.19**	0.3402 ± 0.19

between the generated image and the style cluster, offering an additional robustness check.

4.3 Style-based Recommendation

Collaborative filtering is a widely used approach in recommendation systems, relying on patterns of user-item interactions to predict new preferences without requiring explicit information about the content itself. Rather than analyzing the properties of items, collaborative filtering leverages similarities between users or between items based on historical behavior. A particularly effective category within collaborative filtering is matrix factorization, which models the interactions between users and items by decomposing the observed rating matrix into lower-dimensional latent factors. These latent factors capture underlying user interests and item characteristics, allowing the system to make personalized predictions even in the presence of sparse data. Among matrix factorization techniques, Singular Value Decomposition (SVD) is a popular method that represents user preferences through a combination of global biases, user-specific biases, item-specific biases, and the interaction between learned latent vectors. The parameters of this model are typically optimized using techniques like Stochastic Gradient Descent, enabling accurate and scalable recommendations based on user

behavior patterns.

The sheer volume of styles generated by users presents a significant challenge in navigating the landscape of artistic options available. To address this, we showcase a recommendation system that suggests top styles to users based on their previously generated images. This personalized approach is crucial as it helps users discover styles that align with their tastes and past preferences, simplifying their search among a vast array of choices. We formulate our task as a matrix-factorization-based recommendation, which involves a set of items where users rate items they have interacted with, thus creating a matrix of user-item ratings. In the context of our problem, users are the creators who generate images, and items are the clusters in which generated images are assigned. We calculate for each user u a vector

$$\mathbf{v}_u = \begin{bmatrix} r_1 & r_2 & \dots & r_N \end{bmatrix}$$

where r_i represents the proportion of images that the user has generated within a specific cluster i such that $\sum_{i=1}^{N} r_i = 1$. These vectors create a matrix R where entries r_{ui} denote the rating for user u for cluster i.

We employ the SVD algorithm [13, 63] to obtain a prediction for all

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

where μ is the global average rating, b_u and b_i are the user and item bias terms, respectively, and q_i and p_u are the latent factor vectors for item i and user u, respectively. We minimize the regularized squared error loss and update parameters using Stochastic Gradient Descent [17]. We assess the MAE and RMSE of the predicted ratings against the ground-truth ratings, obtaining a mean RMSE of 0.1425 and a standard deviation of 0.0017, along with a mean

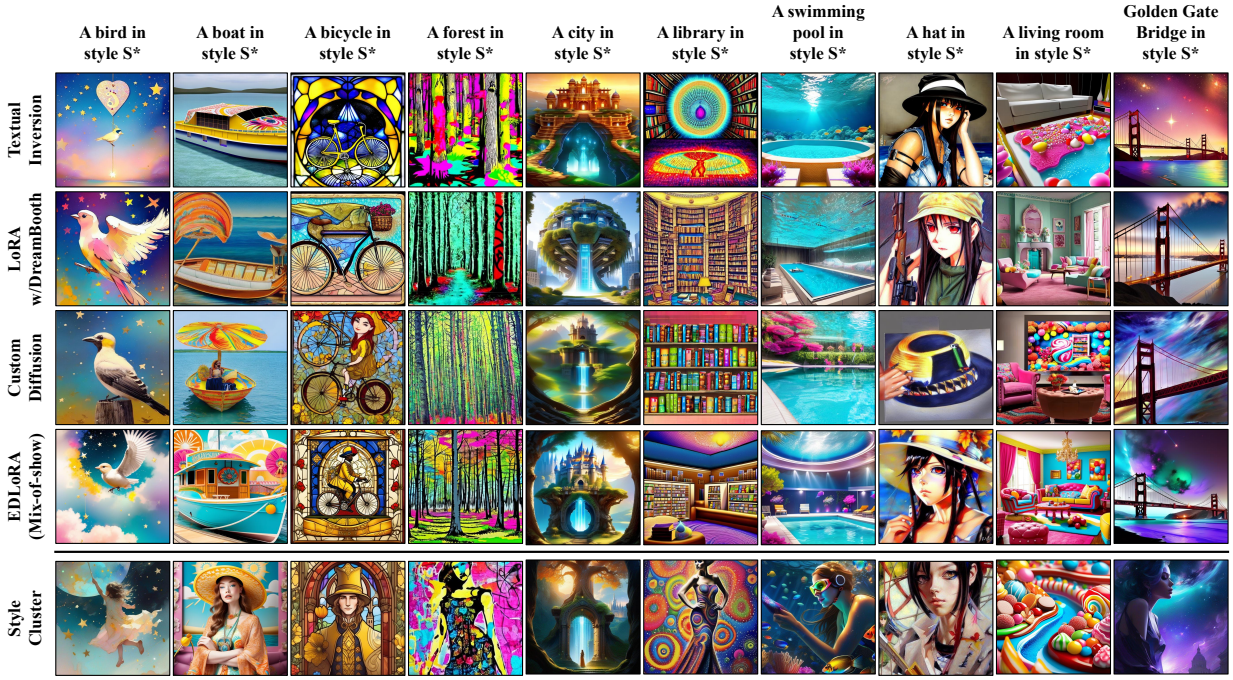


Figure 4.2: Qualitative comparison of various personalization methods on artistic styles on Textual Inversion [14], LoRA w/DreamBooth [26, 50], Custom Diffusion [31], and ED-LoRA [18]. `Style Cluster` (bottom row) illustrates a sample image from the corresponding cluster.

MAE of 0.082 and a standard deviation of 0.001 across all folds. Figure 4.1(d) depicts an example of recommendations for a user based on previously generated styles.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.1585	0.1532	0.1647	0.1678	0.1587	0.1606	0.0051
MAE (testset)	0.0959	0.0938	0.0978	0.0992	0.0970	0.0967	0.0018

Table 4.3: Evaluation metrics across 5-fold cross-validation

4.3.1 Style Atlas for Democratizing Artistic Styles

Since LoRA has emerged as a popular and effective tool for lightweight concept tuning within the community [51], we aim to make this technology easily accessible to users through our platform. Thus, we provide a curated collection of 100 style-specific LoRA models as part of

the Style Atlas platform. These models represent a diverse range of artistic styles discovered through our clustering process, offering users a broad selection of creative options to explore and personalize.

Figure 4.1(c) presents a screenshot of the Style Atlas platform interface, showcasing how users can browse through different available LoRA models. Each model is associated with an example style image, allowing users to visually evaluate and select styles that resonate with their artistic preferences. By simply downloading these LoRA files, users can incorporate specific stylistic elements into their own text-to-image generation workflows without the need for extensive retraining or complex fine-tuning procedures.

To further support ease of use, we also provide a notebook that demonstrates how to load the downloaded LoRA models and apply them for personalized image generation. The notebook offers step-by-step guidance, ensuring that even users with minimal technical background can experiment with new styles and create customized outputs. This functionality aligns with our broader goal of democratizing access to personalized generative tools and empowering a wider range of users to engage creatively with AI.

Our Style Atlas of 100 LoRAs open for download is available at https://stylebreeder.github.io/atlas/ and Fig. 4.3 showcases the website.

4.4 NSFW and Toxic Content

We analyze NSFW content in both images and text prompts using state-of-the-art predictors for images [20] and text prompts [19]. Figure 3.2 (a) shows a comparison of our dataset (`STYLEBREEDER`) with other popular datasets such as LAION [54], Artbench [35], DiffusionDB [69] and TWIGMA [6]. Most of these datasets, particularly those with AI-

[0]NSFW plots for LAION, Artbench, TWIGMA, and DiffusionDB are adopted from [6].

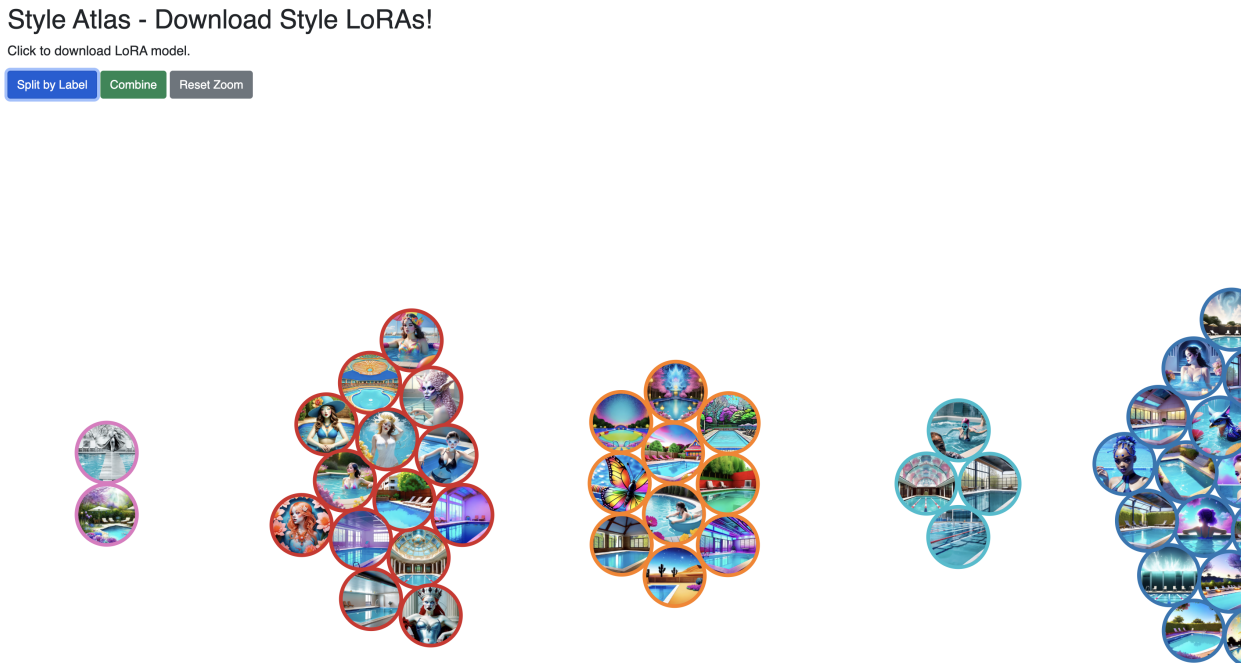


Figure 4.3: Style Atlas Platform

generated images like DiffusionDB, TWIGMA, and ours, contain a substantial amount of potentially NSFW content. A similar observation can be made for text-prompts (see Fig. 3.2 (b)) where we report NSFW, Toxicity, Severe Toxicity, Identity Attack, Obscene, Insult, and Threat scores computed with [19]. This trend correlates with recent studies highlighting a significant increase in NSFW content generation by online communities [47, 53]. For instance, potentially harmful text prompts may involve the names of influential politicians, such as 'Donald Trump' and 'Joe Biden', found in prompts like 'angry Joe Biden screaming, red-faced, steam coming from ears' or 'angry Donald Trump Melania and the judge and police at mcdonalds'. Additionally, our analysis reveals the use of celebrity names in contexts suggesting sexually explicit content that could be construed as nonconsensual pornography. To assist researchers, our dataset includes NSFW text and image scores, along with toxicity-related scores, enabling them to filter these images and determine appropriate thresholds for excluding potentially unsafe data. We also provide a Google form for reporting harmful or

inappropriate images and prompts, as outlined in Section 5.1.

4.5 Temporal Trends in Seasonal Content Generation

Our dataset offers substantial potential for deeper exploration of temporal patterns in user behavior and content generation, particularly in understanding how seasonal variations may influence the types of images and artistic styles that users create throughout the year. By examining changes in prompt content over time, we can begin to uncover how cultural events, holidays, and broader seasonal trends shape user preferences and creative output on generative art platforms.

As an initial example, we conducted a preliminary analysis focusing on the week leading up to Halloween (October 24–31). During this period, we observed a marked increase in the frequency of keywords such as *Halloween*, *scary*, *costume*, and *pumpkin* in user text prompts compared to baseline levels during other times of the year. This clear thematic shift suggests that users adapt their creative prompts to align with relevant cultural and seasonal motifs, leading to noticeable fluctuations in the types of content generated. Further examination of these seasonal trends could yield valuable insights, which we leave as an avenue for future work.

4.6 Identifying Artistic Influences in Text Prompts

On platforms such as Artbreeder, it is common practice for users to include the names of well-known artists within their text prompts as a way to evoke a particular stylistic influence or aesthetic. These artist references serve as shorthand cues, allowing users to guide the text-to-image models toward generating images that emulate specific visual styles associated with

famous creators. To systematically identify and capture these references within our dataset, we employed a widely used Named Entity Recognition (NER) library [66] to extract mentions of artist names directly from the text prompts provided by users.

Following the application of the NER pipeline, we compiled a list of the most frequently referenced artists in our corpus. The top 18 artist names identified through this process are presented in Tab. 4.4, offering insight into the dominant stylistic influences users tend to invoke when generating images on the platform. To make this information readily accessible for further research or analysis, we included an additional column in our released dataset specifically indicating the potential artist names that appear in each corresponding text prompt. This annotation provides a valuable resource for studies related to stylistic conditioning, prompt engineering, or understanding cultural trends in generative art.

Moreover, we recognize the importance of respecting the rights and preferences of artists whose names may be referenced in user-generated prompts. To that end, we have established an opt-out mechanism through a publicly available Google form hosted on our website. Any artist who discovers that their name is included in our dataset and wishes to request removal may do so through this form. We are committed to honoring such requests promptly, as part of our broader efforts to ensure ethical, transparent, and respectful handling of data in the context of AI-generated art.

4.7 Copyright Infringement

One of the primary applications of this large-scale dataset of generated content is to understand how much-generated art is a function of an original artist's work. This has a significant impact on the copyright infringement of T2I models. We apply the method from [42], which performs artistic style classification by determining a particular artist's unique style through

Table 4.4: Top 18 artists used in text-prompts.

Artist Name	Occurrence
Tom Bagshaw	812355
Stanley Artgerm	547422
Greg Rutkowski	521464
Daniel F Gerhartz	430276
WLOP	389215
Charlie Bowater	356740
Atey Ghailan	351338
Andrew Atroshenko	336390
Rossdraws	289541
Edouard Bisson	229375
Alphonse Mucha	211639
Ilya Kuvshinov	206632
Mike Mignola	196128
Pino Daeni	123757
Krenz Cushart	120184
Ismail Inceoglu	107547
Luis Royo	100998
Guweiz	99543

a reference dataset of 372 artists' works. This enables us to recognize if identified styles reappear within images in our dataset. Using this successfully identified all 372 artists supported, resulting in 688K images within our dataset. The top artists discovered using this dataset are highlighted in Tab. 4.5. This demonstrates that our dataset can serve as a valuable resource for scaling up the mentioned method, thereby improving the robustness of artist classification across the 372 artists. Additionally, given the extensive coverage of artists in our dataset, this method could be extended to cover a much broader range of artists, offering valuable insights for addressing copyright infringement issues.

Table 4.5: Top 10 artist styles detected

Artist Name	Accuracy	# Samples in STYLEBREEDER
Alphonse Mucha	42%	209357
Ivan Aivazovsky	39%	50177
Francis Bacon	28%	5565
Claude Monet	28%	8210
Vincent Van Gogh	28%	14671
Hieronymus Bosch	27%	8098
Frank Stella	23%	9254
John William Waterhouse	22%	60683
Egon Schiele	21%	11003
Dan Witz	21%	10964

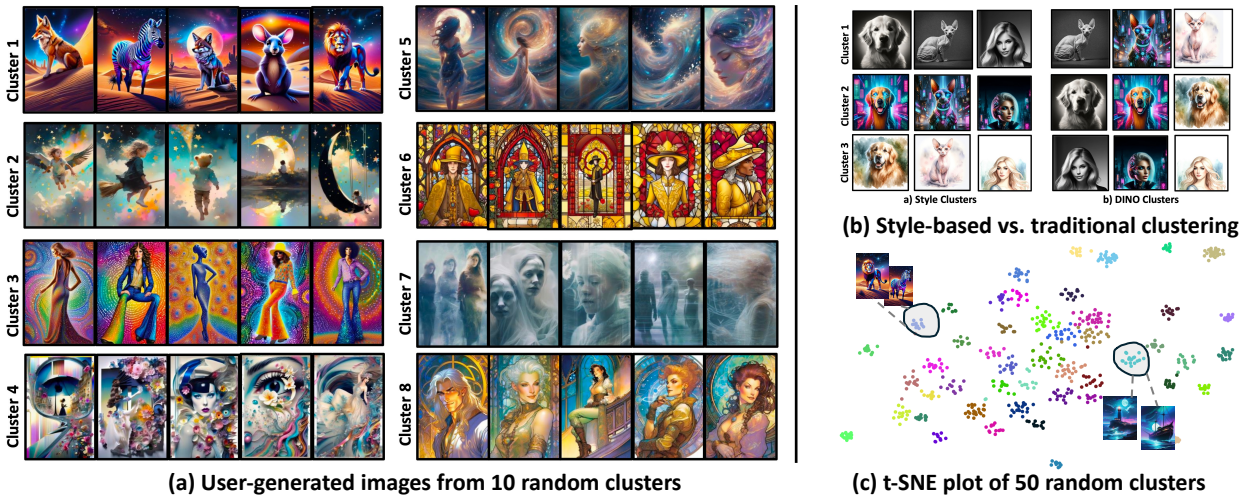


Figure 4.4: (a) User-generated images from 10 random clusters showcasing a diverse range of styles. (b) Sample images from style-based clustering vs. traditional clustering using DINO features show that style-based clustering captures the stylistic content while traditional clustering focuses on objects. (c) Visualization of the clusters, projected into 2D with t-SNE [22] with each cluster represented by a unique color according to their assignments by K-Means++ [1]. This depiction highlights that while many styles are closely related, some distinct styles are noticeably distant from the main clusters.

Chapter 5

Discussion

5.1 Limitations and Societal Impact

While our work significantly advances the integration of AI into creative processes, it also presents certain limitations and societal impacts that warrant careful and critical consideration. It is important to recognize that alongside the benefits of enhancing artistic expression and democratizing access to creative tools, there are inherent challenges that must be acknowledged and addressed.

One of the key limitations lies in the potential for an over-reliance on technology in artistic creation, which could, over time, diminish the perceived value and authenticity of human-driven artistry and creativity. As AI-generated content becomes more widespread and sophisticated, there is a risk that traditional forms of human expression may be undervalued, and the unique role of human artists could be marginalized in certain creative spaces.

Additionally, the use of AI for art generation raises important concerns surrounding copyright, originality, and attribution. When generated styles closely mimic those of existing artists — sometimes without clear, direct attribution — questions arise regarding intellectual property rights and ethical boundaries. This becomes especially significant when considering that many training datasets include artworks without explicit permission from the original creators, further complicating the landscape of ownership and artistic credit.

From a societal perspective, while the tools we propose aim to lower barriers to entry and make artistic creation more accessible to a wider range of users, they also carry the risk of reinforcing existing biases present in the training data. If not carefully monitored, these biases could skew the diversity and representation of the generated artworks, potentially marginalizing underrepresented styles, cultures, or artistic voices. As a result, instead of expanding creative horizons, AI tools could inadvertently narrow them by amplifying dominant aesthetics.

Moreover, as AI-driven creative technologies become increasingly accessible to the public, there is a growing potential for misuse. One particular area of concern is the generation of deceptive imagery or deepfakes, which could erode public trust and further blur the line between authentic and synthetic media. The ability to create realistic but fabricated visuals has broad implications for digital authenticity, social media, journalism, and even democratic processes.

Acknowledging these limitations and societal risks is crucial as we continue to explore and expand the intersection of AI and art. By critically reflecting on these challenges, we can better ensure that the development and deployment of creative AI technologies proceed in ways that are responsible, ethical, and aligned with broader social values.

Chapter 6

Conclusions

This thesis has demonstrated the significant potential of text-to-image diffusion models to explore, capture, and catalog the rich tapestry of user-generated artistic styles on the Artbreeder platform. Through our experiments, we successfully identify and cluster unique, previously uncharted artistic expressions, highlighting the diversity and creativity inherent in user-driven art communities. We further demonstrate the application of these clustered styles by generating personalized images that align with specific aesthetic preferences, showcasing the practical impact of these discoveries for enabling more meaningful and personalized creative experiences.

Additionally, we show that the integration of a personalized recommendation system enhances user engagement by suggesting styles that align closely with individual users' historical preferences. By formulating the problem through a matrix-factorization-based approach and leveraging collaborative filtering techniques, we are able to navigate the vast space of artistic styles and surface relevant suggestions that enrich users' creative journeys. This personalized approach reduces the complexity of discovery and supports deeper artistic exploration, making the platform more accessible and engaging.

Beyond the technical contributions, we also introduce and release the Style Atlas platform, which democratizes access to these innovations. The Style Atlas enables users to browse, download, and experiment with new stylistic models, empowering individuals to incorporate novel artistic elements into their own generative processes. By lowering barriers to entry and

facilitating creative experimentation, the platform aims to foster a more vibrant, inclusive, and dynamic artistic community.

This work not only advances the technological capabilities of AI in the domain of generative art but also contributes meaningfully to the goal of building a more inclusive and diverse artistic ecosystem. By making sophisticated personalization and discovery tools widely available, we help ensure that a broader range of creative voices and aesthetic traditions can be celebrated and explored.

Furthermore, `STYLEBREEDER` and methodologies developed in this thesis open numerous promising avenues for future research. Potential directions include refining the effectiveness of text prompts through iterative prompt optimization, analyzing evolving trends in user-generated art over time, developing recommendation systems that simultaneously consider both image and textual content, building interactive search systems for the generated images based on stylistic or semantic queries, and exploring the concept of explainable creativity [37], which aims to make AI-driven artistic decisions more interpretable and understandable to users. Each of these directions offers opportunities to extend the capabilities of generative AI while continuing to promote creativity, inclusivity, and transparency in the digital art landscape.

Chapter 7

Summary

This thesis explores the use of personalized text-to-image diffusion models to organize, discover, and personalize the broad landscape of user-generated art on platforms such as Artbreeder. We introduce Stylebreeder, a framework that addresses three central tasks: discovering stylistically similar groups through clustering, generating personalized images aligned with individual styles, and recommending new styles based on a user's prior generated content.

To support these objectives, we curate and release a large-scale dataset specifically designed for the study of style discovery and personalization in generative art. The dataset consists of 6.8 million user-generated images along with their associated text prompts, metadata, and additional annotations. For each image, we extracted style-aware features, facilitating high-quality stylistic analysis. We also include cluster assignments for each image based on a large-scale K-Means++[1] clustering process with $k = 10,000$, determined through silhouette score evaluations to optimize internal cluster coherence and external dissimilarity. In addition, we annotate the dataset with potential artist names detected in the prompts using Named Entity Recognition (NER) techniques[66], and offer an opt-out mechanism to address ethical considerations around artist attribution.

Through the clustering process, we reveal both highly individualized artistic signatures as well as broader stylistic trends spanning multiple artists. This foundation enables further experimentation in style-conditioned generation and recommendation. Building on this dis-

covery, we benchmark four state-of-the-art personalization methods—Textual Inversion [14], LoRA with DreamBooth [26, 50], Custom Diffusion [31], and EDLoRA [18]—to generate personalized images corresponding to stylistic clusters. Across multiple evaluation metrics, including CLIP and DINO scores, EDLoRA demonstrates the strongest performance in maintaining both stylistic fidelity and semantic relevance.

To facilitate style discovery for users, we develop a personalized recommendation system based on matrix factorization using collaborative filtering techniques [13, 63]. By modeling user-cluster interaction patterns through user-cluster generation proportions, we predict styles that align with users' preferences, achieving low RMSE and MAE scores and improving the experience of navigating a vast style space.

We make these resources publicly available through the Style Atlas platform, providing access to 100 curated style-specific LoRA models. Users can browse and download models corresponding to different clusters and generate personalized images using a provided notebook for ease of experimentation.

In addition to the primary contributions, our dataset enables a wide range of future research directions, such as refining text prompt effectiveness, analyzing temporal patterns in content generation (e.g., seasonal trends around events like Halloween), building multimodal recommendation systems that jointly consider text and image inputs, developing searchable style repositories, and exploring explainable creativity [37].

Through Stylebreeder, this thesis advances the technological capabilities of generative AI in artistic applications while contributing an openly available dataset and toolset designed to foster a more accessible, diverse, and user-driven creative ecosystem.

Bibliography

[1] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022.

[4] Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*, 2022.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Yiqun Chen and James Y Zou. Twigma: A dataset of ai-generated images with metadata from twitter. *Advances in Neural Information Processing Systems*, 36, 2024.

[7] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference*

on artificial intelligence and statistics, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[8] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[13] Simon Funk. Netflix update: Try this at home, 2006. URL https://sifter.org/simon/journal/20061211.html. Accessed: 2024-06-04.

[14] 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.

[15] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[17] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[18] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023.

[19] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

[20] Laura Hanu and Unitary team. nsfw-detector. Github. https://github.com/LAION-AI/CLIP-based-NSFW-Detector, 2023.

[21] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[22] Geoffrey E. Hinton and Sam T. Roweis. Stochastic neighbor embedding. In *Neural Information Processing Systems*, 2002.

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[27] 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.

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

[29] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021.

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[31] 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.

[32] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[33] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. 2023.

[35] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*, 2022.

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[37] Maria Teresa Llano, Mark d'Inverno, Matthew Yee-King, Jon McCormack, Alon Ilsar, Alison Pease, and Simon Colton. Explainable computational creativity. *arXiv preprint arXiv:2205.05682*, 2022.

[38] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

[39] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11461, 2022.

[40] MidjourneyKaggle. Midjourney user prompts, 2022. URL https://www.kaggle.com/datasets/succinctlyai/midjourney-texttoimage. Accessed: 2024-06-04.

[41] Alexander H. Miller, Will Feng, Dhruva Tirumala, Adam Fisch, Augustus Odena, Vivek Ramavajjala, Joel Z. Leibo, Kelvin Guu andJesse Engel, Jack Clark, Maruan H. Ali,

Nazneen Rajani, Iain J. Dunning, Jacob Andreas, Chris Dyer, Dario Amodei, Jakob Uszkoreit, Douwe Pieksma, Tom Brown, and Ilya Sutskever. Clip: Learning to solve visual tasks by unsupervised learning of language representations. In *International Conference on Machine Learning*, 2020.

[42] Mazda Moayeri, Samyadeep Basu, Sriram Balasubramanian, Priyatham Kattakinda, Atoosa Chengini, Robert Brauneis, and Soheil Feizi. Rethinking artistic copyright infringements in the era of text-to-image generative models. *arXiv preprint arXiv:2404.08030*, 2024.

[43] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

[44] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[45] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

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

[47] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

[48] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[49] 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.

[50] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[51] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2023. URL https://github.com/cloneofsimo/lora.

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[53] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023.

[54] 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. *arXiv preprint arXiv:2210.08402*, 2022.

[55] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li,

and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. *arXiv preprint arXiv:2311.13600*, 2023.

[56] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.

[57] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

[58] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models, 2024.

[59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[61] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021.

[62] Surprise. Surprise, 2019. URL https://surpriselib.com/. Accessed: 2024-06-04.

[63] surprisesvd. Matrix factorization-based algorithms, 2015. URL https://surprise.

readthedocs.io/en/v1.1.1/matrix_factorization.html#unbiased-note. Accessed: 2024-06-04.

[64] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3760–3764. IEEE, 2017.

[65] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://www.aclweb.org/anthology/W03-0419.

[66] Asahi Ushio and Jose Camacho-Collados. T-NER: An all-round python library for transformer-based named entity recognition. In Dimitra Gkatzia and Djamé Seddah, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.7. URL https://aclanthology.org/2021.eacl-demos.7.

[67] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[68] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. Benchmarking deepart detection. *arXiv preprint arXiv:2302.14475*, 2023.

[69] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

[70] Xun Wu, Shaohan Huang, and Furu Wei. Mole: Mixture of lora experts. In *The Twelfth International Conference on Learning Representations*, 2023.

[71] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[72] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4302–4311, 2019.

[73] 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.

[74] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[76] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.