ProCreate, Don't Reproduce! Propulsive Energy Diffusion for Creative Generation

Jack Lu, Ryan Teehan, and Mengye Ren

New York University
{yl11330,rst306,mengye}@nyu.edu
Project Webpage: https://procreate-diffusion.github.io

Abstract. In this paper, we propose ProCreate, a simple and easy-to-implement method to improve sample diversity and creativity of diffusion-based image generative models and to prevent training data reproduction. ProCreate operates on a set of reference images and actively propels the generated image embedding away from the reference embeddings during the generation process. We propose FSCG-8 (Few-Shot Creative Generation 8), a few-shot creative generation dataset on eight different categories—encompassing different concepts, styles, and settings—in which ProCreate achieves the highest sample diversity and fidelity. Furthermore, we show that ProCreate is effective at preventing replicating training data in a large-scale evaluation using training text prompts. Code and FSCG-8 are available at https://github.com/Agentic-Learning-AI-Lab/procreate-diffusion-public.

Keywords: Generative models · few-shot generation · AI-assisted design · data replication

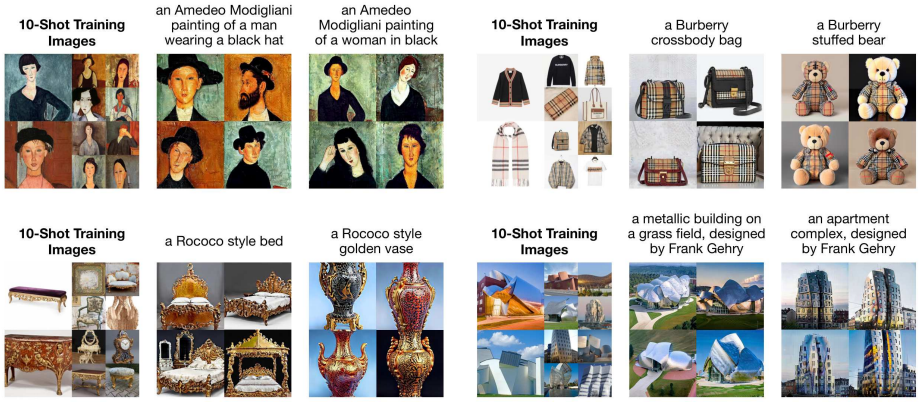


Fig. 1: After fine-tuning a diffusion model on each category of our few-shot dataset FSCG-8, *ProCreate* can significantly improve the diversity and creativity of generations while retaining high image quality and prompt fidelity.

1 Introduction

Imagine a fashion designer attempting to brainstorm a new idea for a clothing line. Looking back on past runway shows that they found particularly stunning and influential, they attempt to draw inspiration from others' work without copying it directly. They want to evoke similar concepts, expressed perhaps in the silhouettes and proportions showcased on the runway, the particular way the fabric drapes on each model's frame, or the drama communicated by a specific cut or color palette. In other words, their goal is to take a reference set of images, which express a unique creative vision from a designer, and draw inspiration from them without direct reproduction. If they attempt to use a generative image model for this creative iteration, however, they would find that the model had either never acquired that ineffable concept before, and thus was unable to produce good examples, or had memorized examples from the reference set during adaptation (e.g. through fine-tuning or tuning new token embeddings [14]), and merely reproduces elements of the design exactly. Our paper aims to address this problem directly, propelling generated images away from those in a reference set while maintaining high-level conceptual inspiration. In that way, we *jointly* address the specific problem of exact replication *and* the general problem of low diversity among generative model samples.

Recent advances in generative image modeling, in particular the development of highly capable, and easier to train, denoising diffusion probabilistic models (DDPMs) [19], have enabled high-quality, complex image generation in both conditional (e.g. class- or text-conditional generation) and unconditional settings [8, 31, 34, 35, 39]. With their widespread adoption, diffusion models are increasingly used in customized settings, where a smaller number of images are used to define a particular domain [4, 54], style [28, 49, 52], subject/concept [14, 25, 36].

Unfortunately, despite their improvements over GANs, diffusion models are similarly prone to training data memorization and a lack of sample diversity [7, 43, 44], which is particularly damaging in light of their use in these customized, low-data settings—generated images are often similar to one another [37]. This behavior is particularly damaging when diffusion models are used to produce art or graphic design materials, where creativity is essential, since the models may replicate potentially copyrighted examples directly [43, 44], opening users and model developers up to potential liability.

To improve sampled image diversity, we propose an energy-based [10, 17, 23, 46, 47] method which applies a propulsive force to latent representations during inference, pushing the generated image away from the images in the reference set. We consider two experimental setups: 1) few-shot creative generation and 2) training data replication prevention. For few-shot generation, we construct a new few-shot image generation dataset called FSCG-8 using collected images across eight different categories. While fine-tuning a diffusion model to a reference set quickly overfits in this low-data setting, ProCreate, by contrast, achieves greater sample diversity while retaining broad conceptual similarity, both being essential components of creative generation. In the training data replication experiments,

we show that ProCreate is effective at preventing pre-trained diffusion models from replicating training data.

In summary, our contributions are as follows:

- To test sample diversity, memorization, and creative generation, we collected a few-shot generation dataset, FSCG-8, spanning across domains such as paintings, architecture, furniture, fashion, cartoon characters, etc.
- We propose ProCreate, a simple and easy-to-implement component that allows diffusion models to generate diverse and creative images using concepts from a reference set, without direct reproduction.
- In few-shot generation, ProCreate is found to have better sample diversity than prior arts, while maintaining a high similarity to the reference set.
- ProCreate addresses the training data replication issue with a significantly lower chance of replicating training images than pre-trained diffusion models.

2 Related Work

In this section, we review related areas of work in guided diffusion models, few-shot image generation, sample diversity, and data replication issues.

Guided Diffusion Models. Conditioning provides a powerful and flexible way to guide diffusion model generation using text [35, 40], images [20, 29, 38], videos [18, 22], layout [53], or even scene graphs [11, 50] and point clouds [33, 51]. Recent works have explored classifier-free guidance and classifier guidance as two methods for class- or text-conditional image generation [8, 21]. Classifier-free guidance requires training a new DDPM that accepts an additional conditioning input, but achieves strong performance for conditional image generation tasks, such as conditioning on class [8], prompts or regions of the original image [31], among others. On the other hand, classifier guidance avoids re-training by guiding a pre-trained model using a classifier to guide the sampling process. Recent work applied classifier guidance across a variety of conditioning goals without training new classifiers [3]. DOODL [48] improves the performance of off-the-shell classifier guidance by performing backpropagation through the entire diffusion inference process to optimize the initial noise.

Few-Shot Image Generation. Diffusion models can be adapted to a few-shot setting [15], including for customization and personalization [14, 36]. Giannone *et al*. [15] developed a method for few-shot adaptation of diffusion models, but they focused on CIFAR-100 [24], which only contains 32×32 images. In contrast, we collect our own few-shot generation dataset with 512×512 images, where we choose each class to be practical for designers and for the images of each class to contain more conceptual similarities (e.g., Burberry design) than simply belonging to the same semantic category (e.g., horse). Standard fine-tuning, Dreambooth [36], and Textual Inversion [14] all allow for customizing diffusion models. The first two methods require fine-tuning an entire diffusion model to produce images with a consistent subject, which makes it prone to overfitting. Textual Inversion learns a new token for the concept it wants to replicate, which can have a more limited capacity for representing novel concepts from new training images. Since ProCreate is applicable to any diffusion model, it can easily be

applied on top of either of these approaches. At the same time, neither method handles issues with memorization and sample diversity, whereas ProCreate can adapt to the concept(s) in the few-shot reference set without direct reproduction or low sample diversity.

Sample Diversity. CADS [37] is recently proposed as a method to address the sample diversity problem in diffusion models. CADS uses annealed conditioning during inference, allowing them to balance the quality and diversity of samples from the model. In contrast, ProCreate guides the generated sample to be different from the set of training images, achieving even better sample diversity than CADS and simultaneously avoiding reproducing potentially copyrighted images in the training set. Sehwag *et al.* [42] introduces a framework that involves sampling from low-density regions of the data manifold to avoid reproducing training images and generating novel samples. However, their method operates in pixel space rather than in latent space like ours, making it difficult to apply to latent diffusion models. Finally, we explore the low-data few-shot setting where one wants to not only avoid reproduction but also draw inspiration from the limited training images, while CADS and Sehwag *et al.* [42] do not.

Data Replication. We also draw on prior work studying data replication in generative models, particularly diffusion models. Carlini *et al.* [6] constructed a pipeline that is used to extract thousands of training examples from state-of-the-art diffusion models, Somepalli *et al.* [43] studied the rate at which diffusion models' inference outputs replicate their training data at various data sizes, and Somepalli *et al.* [44] proposed de-duplicating images in the training set and randomizing/augmenting captions to reduce the rate of output replications. Similar issues were studied in GANs both theoretically [30] and empirically [1,2]. In contrast to Somepalli *et al.* [44], ProCreate addresses the issue of diffusion output replication more directly by guiding each generation away from a set of images (e.g., all training images) we want to avoid generating.

3 Background

In this section, we will cover the basics of diffusion denoising probabilistic models (DDPM) [19], denoising diffusion implicit models (DDIM) [45], and classifier guidance [8]. Given samples from a data distribution $q(\boldsymbol{x}_0)$ of images, we can use diffusion models to learn a model distribution $p_\theta(\boldsymbol{x}_0)$ that approximates $q(\boldsymbol{x}_0)$ and can be sampled from. To sample a new image, diffusion denoising probabilistic models (DDPMs) [19] start with a sample of Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then repeatedly denoise $\boldsymbol{x}_t$ into $\boldsymbol{x}_{t-1}$ where t decrements from T to 1. In this paper, we use a more performant sampling method than the default DDPM approach, termed denoising diffusion implicit models (DDIM) [45], which was shown to beat GAN models with only 25 sampling steps [8]. At each denoising step with a noisy image $\boldsymbol{x}_t$, the DDIM sampling process first makes a one-step prediction of $\hat{\boldsymbol{x}}_0$ with Equation 1, then it denoises $\boldsymbol{x}_t$ into $\hat{\boldsymbol{x}}_{t-1}$ by Equation 2. $\boldsymbol{\epsilon}_\theta$ represents the learned diffusion model with weights θ and α_t are

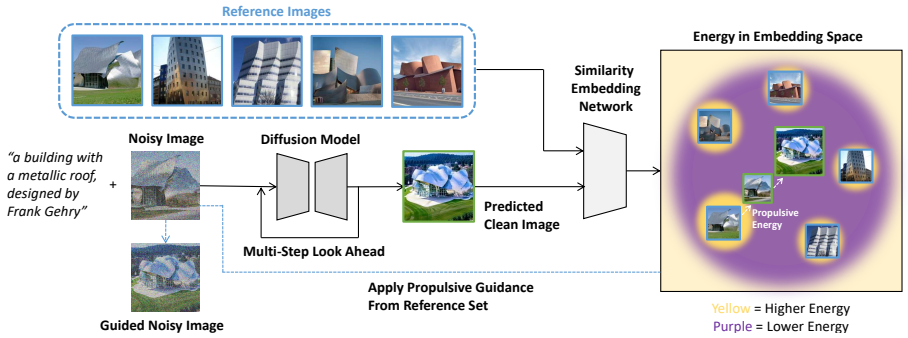


Fig. 2: Overview of our approach. At each denoising step, ProCreate applies gradient guidance that maximizes the distances between the generated clean image and the reference images in the embedding space of a similarity embedding network. In the embedding space, the noisy image is propelled away from its closest reference image.

scalar parameters that define the diffusion noise schedule.

$$\hat{\boldsymbol{x}}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)}{\sqrt{\alpha_t}}, \tag{1}$$

$$DDIM(\boldsymbol{x}_t, t) = \sqrt{\alpha_{t-1}}\hat{\boldsymbol{x}}_0 + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t). \tag{2}$$

To generate samples from diffusion models that are conditioned on ImageNet class labels, Dhariwal *et al.* [8] introduced classifier guidance. They used the gradient of a noise-aware ImageNet classifier $f_\phi(y|\boldsymbol{x}_t)$ to perturb the noise prediction of each denoising step. Following the formulation in Bansal *et al.* [3], classifier-guidance of class label c is applied by modifying the diffusion process with an additive gradient term:

$$\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) + \sqrt{1-\alpha_{t-1}}\nabla_{\boldsymbol{x}_t} l_{ce}(c, f_{cl}(\boldsymbol{x}_t)), \tag{3}$$

where l_{ce} is the cross-entropy loss and f_{cl} is a noise-aware classifier, then replacing $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ with $\boldsymbol{\epsilon}'$ in Equation 2. Furthermore, Ho and Salimans [21] formulated diffusion models as score-matching models and showed that Equation 3 works by lowering the energy of data for which the classifier f_{cl} predicts to be of class c with high likelihood. Note that here, lower energy corresponds to a higher probability of being sampled. In the next section, we will use the notations and basic concepts and notations from here to develop ProCreate.

4 ProCreate: Propulsive Energy Diffusion for Creative Generation

In this section, we introduce the mathematical formulation of ProCreate and the techniques we use to improve ProCreate's performance. The core idea behind ProCreate is to guide the generation away from images in the reference set using the gradient of an energy function computed with that set. As illustrated in Figure 2, at each denoising step t with the current noisy image $\boldsymbol{x}_t$, ProCreate

uses the diffusion model to predict a clean image $\boldsymbol{x}_0$, compute distances from $\boldsymbol{x}_0$ to each reference image, then finally updates $\boldsymbol{x}_t$ with a guidance gradient to maximize the distance between $\boldsymbol{x}_0$ with the closest reference image to it. With the guided noisy version of $\boldsymbol{x}_t$, ProCreate can resume the normal diffusion model sampling process to denoise it to $\boldsymbol{x}_{t-1}$.

Energy Formulation. Mathematically, we start with a reference set of m images: $X^r = \{\boldsymbol{x}_1^r, \ldots, \boldsymbol{x}_m^r\}$. Typically, they are either in the fine-tuning set for few-shot generation or the pre-training set. To mitigate the tendency of models to replicate these images, we guide the denoising process away using a gradient step from the log energy function that detects similarity among these images:

$$\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) + \sqrt{1-\alpha_{t-1}}\nabla_{\boldsymbol{x}_t} g(\boldsymbol{x}_t, \{\boldsymbol{x}_1^r, \ldots, \boldsymbol{x}_m^r\}), \tag{4}$$

where the g is defined as the log energy, which is the similarity between the predicted clean image $\hat{\boldsymbol{x}}_0$ and the closest reference image in the embedding space by using an embedding function f:

$$g(\boldsymbol{x}_t, \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m^r\}) = \gamma \max_{i=1\ldots m} s(f(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)), f(\boldsymbol{x}_i^r)). \tag{5}$$

In practice, we use the cosine similarity for s, and we use DreamSim [13] as our embedding function f since the DreamSim network is already pre-trained on detecting similar replicated images. γ controls the strength of our energy function guidance. Note that the predicted clean image $\hat{\boldsymbol{x}}_0$ is a function of $\boldsymbol{x}_t$ that is predicted by Multi-Step Look Ahead, which we will explain below.

Multi-Step Look Ahead Prediction. Referring to Section 3, $\hat{\boldsymbol{x}}_0$ can be predicted in one-step by Equation 1. However, the one-step prediction of $\hat{x}_0$ at denoising steps when t is large can be very inaccurate. Following DOODL [48], to improve the quality of our estimate for $\hat{x}_0$ at each denoising step and in turn the quality of guidance, we perform DDIM diffusion n_{step} times when predicting $\hat{x}_0$. Specifically, at denoising step t with a noisy sample x_t, the Multi-Step Look Ahead prediction performs the following operations:

$\boldsymbol{x} \leftarrow \boldsymbol{x}_t$
`for` t' `in` n_{step} `timesteps evenly spaced from` t `to 1:`
    $\boldsymbol{x} \leftarrow DDIM(\boldsymbol{x}, t')$
$\hat{\boldsymbol{x}}_0 \leftarrow \boldsymbol{x}$

Dynamically Growing Reference Set. To further improve the diversity of our generated samples, we add newly generated samples to the reference set after each batch $X^b = \{\boldsymbol{x}_0^{b,i}\}$ is generated: $X^r \leftarrow X^r \cup X^b$. Therefore, the reference set continuously grows as we generate more samples, and new samples are guided away from the old ones to prevent diffusion models from generating similar images to previous ones.

5 Few-Shot Creative Generation

Given a diffusion model checkpoint that is fine-tuned on limited data, our goal is to improve the diversity of its generated samples while maintaining sample quality and conceptual similarity to the reference set. In this section, we

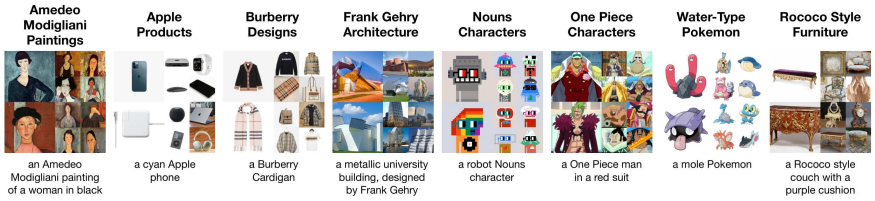


Fig. 3: Samples of the FSCG-8 dataset. We provide 10 samples from each of the 8 categories. For each category, an example caption for its top left image is provided.

describe our experiments on few-shot creative generation, comparing the samples generated from the default DDIM, CADS, and ProCreate sampling methods both qualitatively and quantitatively.

5.1 Our Dataset: FSCG-8

Our goal is to improve the performance of diffusion models on any few-shot training set with general shared properties, which can be subject, style, texture, and more. Diffusion models have been applied to a variety of few-shot learning tasks, Textual Inversion [14] and Dreambooth [36] learn specific subjects, DomainStudio [54] performs domain adaptation, and Specialist Diffusion [28] learns styles, but none of these works curate datasets that share general properties. Therefore, we curate a new dataset FSCG-8 (Few-Shot Creative Generation 8), containing 8 categories of images with 50 prompt-image pairs in each. Samples of the dataset are shown in Figure 3. Each category contains images that share some properties, including the style of Amedeo Modigliani's paintings, the abstract design and texture of Burberry's apparel, and the twisting geometric properties of Frank Gehry's architecture. We manually collect all images from the public domain on the Internet. For each category, we also design the prompts to be simple so that when given a validation prompt to a model, it has a large creative space of images to explore that all follow the prompt.

5.2 Experiment Setup

Fine-tuning. We split each dataset in FSCG-8 into 10 training images and 40 validation images then fine-tune a pre-trained Stable Diffusion v1-5 checkpoint with batch size 8 and learning rate 5×10^{-6} with no learning rate warm-up. We perform fine-tuning with two different training methods: standard fine-tuning for 2000 iterations and DreamBooth fine-tuning for 6000 iterations [36]. DreamBooth fine-tuning is performed with prior preservation and we substitute the [V] token for class nouns "Amedeo", "Apple", "Burberry", "Frank Gehry", "Nouns", "One Piece", "Pokemon", and "Rococo" for each dataset in the order of Figure 3 respectively. For prior preservation, we set its loss weight to 0.5, generate 100 training images from prompts with the class noun removed, and split each batch such that half of the captions contain class nouns and the other half do not.

Inference. For each trained checkpoint, we compare the quality and diversity of sampled images with DDIM, CADS, and ProCreate. For each evaluation run, we generate 40 samples from the 40 prompts in the validation split. We set the number of inference steps to 50 for all sampling methods. We tune hyperparam-

eters t_2, s for CADS and tune γ, gradient norm clipping (on the gradient of the energy function) for ProCreate. For ProCreate, we set the reference set to the training set and $n_{step} = 5$ for Multi-Step Look Ahead.



Fig. 4: Qualitative comparison between DDIM, CADS, and ProCreate for few-shot creative generation on FSCG-8 with standard fine-tuning. For each sampling method, we show two prompts and four generated samples for each prompt. In addition, we match each sample from ProCreate with its closest training image based on the SSCD score [32] between the matched pair.

Evaluation Metrics. We use FID [16] and KID [5] scores to measure how well our generated sample distribution matches that of real images, Precision and Recall [26] to measure the quality and diversity/coverage respectively, and following CADS, we use the Mean Similarity Score (MSS) and the Vendi score [12, 37] to measure diversity of the generated samples. Specifically, we evaluate FID and KID with 64 feature dimensions and set $k = 5$ for Precision and Recall. To ensure that our generated images follow their input prompts, we also use CLIP to compute the cosine similarity between the prompt embedding and the generated image embedding, calling it Prompt Fidelity.

5.3 Results

Creative Generation Visualization. We perform standard fine-tuning on a separate pre-trained stable diffusion for each dataset, then use the same checkpoint at 2000 iterations to perform DDIM, CADS, and ProCreate sampling. Figure 4 shows two example input prompts and their respective outputs from each sampling method. We observe that while the images produced by DDIM sampling do share similar properties and details as their 10-shot training images in Figure 3, each image grid of 2-by-2 samples is very perceptually similar, lacking diversity. While CADS' samples can achieve better sample diversity than the default DDIM sampling, ProCreate generates much more diverse images that balance between including shared properties in the training images and following the prompt guidance. Consider the generated samples of the prompt "an Amedeo Modigliani painting of a girl in blue", DDIM's 4 samples are perceptually similar and only one of the CADS outputs significantly varies in posture and cloth color, while ProCreate outputs contain women with diverse faces, hair color, background, and clothing. For the Top-1 SSCD [32] columns, we select the closest training image to each ProCreate sample based on the SSCD score, where a higher score suggests a higher likelihood of training data replication. Since all ProCreate samples look significantly different from their matched training images, ProCreate does not replicate its training images.

Quantitative Evaluation with Standard Fine-Tuning. In Table 1, we use the metrics in Section 5.2 to evaluate the same set of standard fine-tuning checkpoints used for qualitative results. Due to the small dataset size, we evaluate each checkpoint and method 5 times with random train/validation splits to obtain the mean and standard deviation values for each metric. On close to all datasets of FSCG-8, not only does ProCreate achieve significantly better performance than DDIM and CADS in all diversity-focused metrics (Recall, MSS, Vendi), ProCreate is also superior in metrics that measure both quality and diversity (FID, KID). ProCreate being competitive with other sampling methods in Precision and Prompt Fidelity shows that it can generate diverse samples while preserving high output quality and fidelity.

Quantitative Comparison with DreamBooth Fine-Tuning. In addition to standard fine-tuning, we can also apply ProCreate on DreamBooth [36], where a DreamBooth model is fine-tuned on each category of the few-shot dataset. Table 2 shows the evaluation results for checkpoints obtained from DreamBooth training. ProCreate again achieves significantly better performance on FID, KID,

Table 1: Quantitative comparison between DDIM, CADS, and ProCreate samples from standard fine-tuning checkpoints for 10-shot learning on various generative modeling metrics. We show the mean and standard deviation values over 5 repeated runs in each cell.

Subset	Method	FID ↓	KID ↓	Precision ↑	Recall ↑	MSS ↓	Vendi ↑	Prompt Fid. ↑
Amedeo	DDIM	16.71 ± 0.39	2.39 ± 0.10	**0.75 ± 0.09**	0.36 ± 0.06	0.34 ± 0.01	14.89 ± 0.69	**0.35 ± 0.00**
	CADS	12.83 ± 0.48	2.29 ± 0.12	0.73 ± 0.05	0.39 ± 0.02	0.34 ± 0.01	15.17 ± 0.40	**0.35 ± 0.00**
	ProCreate	**9.28 ± 0.39**	**1.59 ± 0.25**	0.57 ± 0.05	**0.65 ± 0.05**	**0.26 ± 0.01**	**22.20 ± 0.55**	**0.35 ± 0.00**
Apple	DDIM	48.51 ± 3.84	1.37 ± 0.13	**0.51 ± 0.03**	0.75 ± 0.07	0.17 ± 0.01	25.13 ± 0.69	**0.30 ± 0.00**
	CADS	38.88 ± 1.47	1.17 ± 0.22	0.43 ± 0.04	0.79 ± 0.01	0.17 ± 0.01	26.34 ± 0.87	**0.30 ± 0.00**
	ProCreate	**24.59 ± 0.14**	**0.62 ± 0.10**	0.45 ± 0.06	**0.81 ± 0.02**	**0.12 ± 0.00**	**32.33 ± 0.37**	0.29 ± 0.00
Burberry	DDIM	35.11 ± 2.60	4.06 ± 0.48	**0.69 ± 0.06**	0.71 ± 0.05	0.18 ± 0.00	26.15 ± 0.56	**0.34 ± 0.00**
	CADS	37.71 ± 2.09	4.43 ± 0.37	**0.69 ± 0.07**	0.74 ± 0.03	0.19 ± 0.01	25.46 ± 0.57	**0.34 ± 0.00**
	ProCreate	**14.10 ± 1.64**	**0.72 ± 0.11**	0.66 ± 0.06	**0.97 ± 0.02**	**0.10 ± 0.00**	**33.51 ± 0.36**	0.33 ± 0.00
Frank	DDIM	6.36 ± 0.37	0.37 ± 0.06	**0.99 ± 0.01**	0.58 ± 0.05	0.17 ± 0.01	26.21 ± 0.57	**0.32 ± 0.00**
	CADS	4.65 ± 0.32	0.35 ± 0.06	0.98 ± 0.02	0.59 ± 0.06	0.18 ± 0.01	26.49 ± 0.23	**0.32 ± 0.00**
	ProCreate	**3.20 ± 0.24**	**0.20 ± 0.02**	0.94 ± 0.04	**0.67 ± 0.03**	**0.16 ± 0.01**	**29.70 ± 0.58**	**0.32 ± 0.01**
Nouns	DDIM	2.83 ± 0.04	0.04 ± 0.00	**0.12 ± 0.04**	0.81 ± 0.11	0.47 ± 0.02	11.50 ± 0.51	**0.25 ± 0.00**
	CADS	**2.54 ± 0.06**	**0.03 ± 0.00**	0.11 ± 0.03	0.82 ± 0.04	0.47 ± 0.01	11.50 ± 0.23	**0.25 ± 0.00**
	ProCreate	2.72 ± 0.04	**0.03 ± 0.00**	**0.12 ± 0.03**	**0.91 ± 0.06**	**0.42 ± 0.00**	**12.07 ± 0.17**	**0.25 ± 0.00**
Onepiece	DDIM	6.13 ± 0.50	0.67 ± 0.05	0.71 ± 0.04	0.64 ± 0.06	0.26 ± 0.01	23.73 ± 0.29	**0.30 ± 0.00**
	CADS	7.13 ± 0.22	0.81 ± 0.11	0.71 ± 0.04	0.62 ± 0.02	0.27 ± 0.01	22.95 ± 0.40	**0.30 ± 0.00**
	ProCreate	**4.84 ± 0.24**	**0.44 ± 0.07**	**0.72 ± 0.05**	**0.67 ± 0.01**	**0.25 ± 0.01**	**25.80 ± 0.27**	**0.30 ± 0.00**
Pokemon	DDIM	14.29 ± 0.94	0.28 ± 0.02	0.44 ± 0.07	0.77 ± 0.10	0.30 ± 0.01	21.46 ± 0.60	**0.32 ± 0.00**
	CADS	16.57 ± 0.39	0.28 ± 0.02	**0.48 ± 0.10**	0.75 ± 0.03	0.29 ± 0.01	22.29 ± 0.35	**0.32 ± 0.00**
	ProCreate	**11.38 ± 0.76**	**0.27 ± 0.01**	0.44 ± 0.07	**0.84 ± 0.04**	**0.28 ± 0.01**	**22.51 ± 0.37**	0.31 ± 0.00
Rococo	DDIM	26.47 ± 1.17	5.64 ± 0.41	**0.95 ± 0.01**	0.83 ± 0.03	0.16 ± 0.00	22.68 ± 0.28	**0.33 ± 0.00**
	CADS	30.10 ± 1.35	6.63 ± 0.43	0.92 ± 0.02	0.78 ± 0.04	0.17 ± 0.00	22.22 ± 0.28	**0.33 ± 0.00**
	ProCreate	**17.96 ± 1.44**	**3.26 ± 0.36**	**0.95 ± 0.03**	**0.89 ± 0.02**	**0.10 ± 0.01**	**23.43 ± 1.15**	0.32 ± 0.00

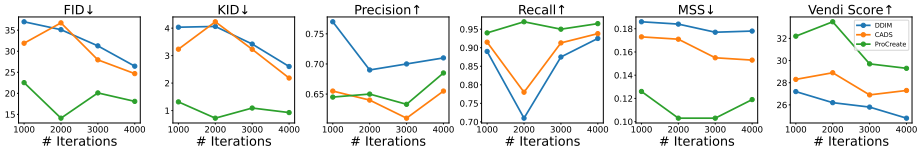


Fig. 5: Compare DDIM, CADS, and ProCreate's performance under different numbers of fine-tuning iterations.

Recall, MSS, and Vendi Score on almost all datasets while remaining competitive with our baselines in Precision and Prompt Fidelity. The results of this experiment show that ProCreate improves the performance of a state-of-the-art few-shot learning method and gives strong evidence for ProCreate's general ability to improve any checkpoint fine-tuned on limited samples.

5.4 Ablation Experiments

Fine-Tuning Steps. We evaluate the baselines and ProCreate on checkpoints at training iterations 1k, 2k, 3k, and 4k of the standard fine-tuning run on the "Burberry Designs" dataset. Figure 5 shows that ProCreate consistently improves FID, KID, Recall, MSS, and Vendi Score and improves the trade-off in Precision in comparison to CADS.

Number of Inference Steps. We evaluate the baselines and ProCreate on the number of diffusion inference steps from 20 to 50. ProCreate again shows consistent improvements in all metrics except Precision while achieving a better trade-off in Precision in comparison to CADS.

Table 2: Quantitative comparison between DDIM, CADS, and ProCreate samples from DreamBooth fine-tuning checkpoints for 10-shot learning on various generative modeling metrics. We show the mean and standard deviation values over 5 repeated runs in each cell.

Subset	Method	FID ↓	KID ↓	Precision ↑	Recall ↑	MSS ↓	Vendi ↑	Prompt Fid. ↑
Amedeo	DDIM	17.38 ± 0.40	2.96 ± 0.13	0.48 ± 0.04	0.63 ± 0.15	0.28 ± 0.00	14.12 ± 0.37	**0.32 ± 0.00**
	CADS	16.19 ± 0.56	2.56 ± 0.17	**0.53 ± 0.06**	0.65 ± 0.07	0.27 ± 0.00	14.37 ± 0.40	**0.32 ± 0.00**
	ProCreate	**8.88 ± 0.17**	**1.19 ± 0.27**	0.45 ± 0.05	**0.80 ± 0.01**	**0.20 ± 0.01**	**25.49 ± 0.55**	0.31 ± 0.00
Apple	DDIM	29.71 ± 0.63	0.70 ± 0.07	0.39 ± 0.02	0.71 ± 0.05	0.22 ± 0.01	17.38 ± 1.04	**0.27 ± 0.00**
	CADS	32.52 ± 2.21	0.84 ± 0.12	0.35 ± 0.06	0.71 ± 0.02	0.21 ± 0.01	18.82 ± 1.39	**0.27 ± 0.00**
	ProCreate	**20.12 ± 2.44**	**0.27 ± 0.05**	**0.40 ± 0.03**	**0.84 ± 0.04**	**0.15 ± 0.00**	**28.23 ± 0.83**	**0.27 ± 0.00**
Burberry	DDIM	21.89 ± 2.24	2.28 ± 0.33	0.56 ± 0.06	0.78 ± 0.07	0.20 ± 0.01	20.85 ± 0.57	**0.28 ± 0.00**
	CADS	19.99 ± 1.53	2.10 ± 0.36	0.54 ± 0.07	0.81 ± 0.12	0.21 ± 0.00	20.26 ± 0.40	0.27 ± 0.00
	ProCreate	**10.18 ± 1.03**	**0.37 ± 0.06**	**0.59 ± 0.06**	**0.92 ± 0.05**	**0.12 ± 0.00**	**28.64 ± 0.85**	**0.28 ± 0.00**
Frank	DDIM	3.88 ± 0.14	0.34 ± 0.02	**0.71 ± 0.02**	0.65 ± 0.01	0.20 ± 0.00	18.66 ± 0.30	**0.26 ± 0.00**
	CADS	3.42 ± 0.27	0.31 ± 0.05	0.70 ± 0.02	0.70 ± 0.02	0.20 ± 0.00	18.62 ± 0.28	**0.26 ± 0.00**
	ProCreate	**2.38 ± 0.18**	**0.09 ± 0.01**	0.57 ± 0.03	**0.78 ± 0.02**	**0.14 ± 0.01**	**26.99 ± 1.41**	**0.26 ± 0.00**
Nouns	DDIM	0.65 ± 0.02	0.02 ± 0.00	**0.77 ± 0.07**	0.39 ± 0.03	0.61 ± 0.01	6.50 ± 0.17	**0.25 ± 0.00**
	CADS	0.66 ± 0.02	0.04 ± 0.00	0.76 ± 0.03	0.38 ± 0.03	0.61 ± 0.00	6.41 ± 0.09	**0.25 ± 0.00**
	ProCreate	**0.62 ± 0.01**	**0.01 ± 0.00**	0.76 ± 0.04	**0.43 ± 0.02**	**0.56 ± 0.01**	**7.24 ± 0.18**	**0.25 ± 0.00**
Onepiece	DDIM	5.16 ± 0.26	0.32 ± 0.03	**0.36 ± 0.01**	0.71 ± 0.02	0.23 ± 0.00	18.72 ± 0.19	**0.25 ± 0.00**
	CADS	5.33 ± 0.30	0.36 ± 0.04	0.26 ± 0.02	0.70 ± 0.02	0.23 ± 0.00	18.78 ± 0.41	**0.25 ± 0.00**
	ProCreate	**4.34 ± 0.19**	**0.13 ± 0.03**	0.28 ± 0.02	**0.86 ± 0.02**	**0.20 ± 0.00**	**19.98 ± 0.38**	**0.25 ± 0.00**
Pokemon	DDIM	**8.13 ± 0.82**	**0.13 ± 0.03**	**0.46 ± 0.04**	0.85 ± 0.03	0.25 ± 0.00	21.08 ± 0.68	**0.25 ± 0.00**
	CADS	13.51 ± 0.89	0.25 ± 0.03	0.38 ± 0.03	0.88 ± 0.05	0.21 ± 0.01	24.32 ± 0.75	**0.25 ± 0.00**
	ProCreate	11.67 ± 0.47	0.20 ± 0.03	0.41 ± 0.06	**0.92 ± 0.04**	**0.17 ± 0.02**	**27.70 ± 1.77**	**0.25 ± 0.00**
Rococo	DDIM	13.57 ± 1.09	2.00 ± 0.28	**0.93 ± 0.00**	0.77 ± 0.02	0.17 ± 0.00	12.10 ± 0.96	0.24 ± 0.00
	CADS	16.27 ± 1.26	2.59 ± 0.30	0.83 ± 0.00	0.80 ± 0.01	0.17 ± 0.00	12.15 ± 0.40	0.24 ± 0.00
	ProCreate	**9.35 ± 1.06**	**1.12 ± 0.19**	0.88 ± 0.05	**0.85 ± 0.02**	**0.12 ± 0.01**	**23.07 ± 1.95**	**0.25 ± 0.00**

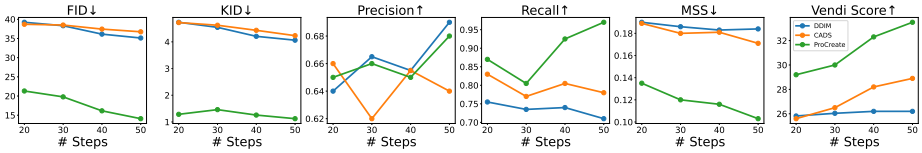


Fig. 6: Compare DDIM, CADS, and ProCreate's performance under different numbers of inference steps.

Additional Ablation Results. We also experimented with varying the diffusion samplers (e.g., DDPM, PNDM [27]), the number of training samples (e.g., 5-shot learning, 25-shot learning), and varying the number of steps for Multi-Step Look Ahead prediction. These results are included in the Supplementary Materials.

6 Training Data Replication Prevention

With the surging interest in generative AI in recent years, people use image generation models for a variety of entertainment or business purposes. However, recent studies show that even large-scale diffusion models are prone to replicating data inside its training set [43, 44], raising privacy and copyright concerns. To prevent large-scale diffusion models from replicating their training data, we follow the setup of Somepalli *et al.* [43] for this experiment to test ProCreate's ability to guide generations of pre-trained Stable Diffusion models away from their LAION [41] training dataset.

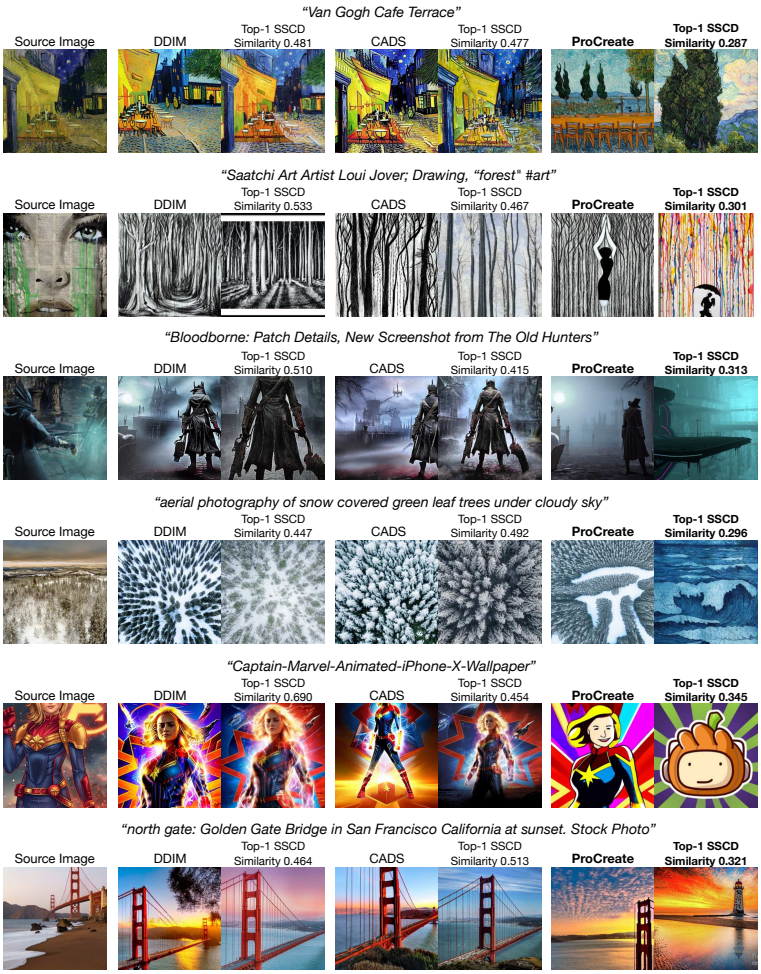


Fig. 7: Qualitative comparison between DDIM, CADS, and ProCreate for replication prevention on LAION12M.

6.1 Experiment Setup

Model and Dataset. We use the frozen pre-trained Stable Diffusion v1-5 checkpoint that is pre-trained on the LAION dataset containing 2B prompt-image pairs. To scale the compute within our limitation, we follow Somepalli *et al.* [43] by setting the reference set to the smaller LAION Aesthetics v2 6+ dataset with 12M caption-image pairs and is a subset of images used in the last stage of Stable Diffusion v1-5 fine-tuning, namely LAION12M. Using a random subset of prompts from LAION12M, we generate 9k samples with Stable Diffusion v1-5 and search each sample's matching LAION12M image with the highest Top-1 SSCD score.

Inference Implementation. We perform 50 inference steps and set ProCreate's Multi-Step Look Ahead n_{step} to 5. To perform inference efficiently, before gen-

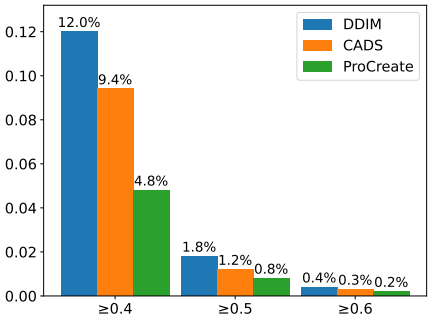


Fig. 8: Comparison of DDIM, CADS, and ProCreate on the percentage of generated 9k images over a Top-1 SSCD threshold.

	FID	**KID**
DDIM	1.38	0.051
CADS	1.60	0.065
ProCreate	**1.14**	**0.038**

Table 3: Comparison of DDIM, CADS, and ProCreate on FID and KID.

n_{step}	**Inference Time (s/sample)**
0	3.1
1	18.5
3	28.3
5	37.0

Table 4: MSLA n_{step}'s effect on ProCreate's inference time with NVIDIA A100.

erating each sample, we filter LAION12M to 10k images that have the most similar captions to the sample caption based on their CLIP embeddings' cosine similarity. We use the FAISS [9] library to speed up vector similarity searches. We again compare ProCreate to DDIM and CADS.

Evaluation Metrics. To evaluate how well ProCreate prevents data replication, we compute the percentages of Top-1 SSCD scores over the thresholds 0.4, 0.5, and 0.6. To ensure that the generated images are still within the distribution of LAION12M images, we also compare them with FID and KID.

6.2 Results

Image Generation Visualization. Figure 7 shows examples of when the pre-trained Stable Diffusion model generates images that are both perceptually similar to their matched images in LAION12M and high in Top-1 SSCD score. While CADS sampling reduces the Top-1 SSCD score in most cases, the "Van Gogh Cafe Terrace" and "Golden Bridge" examples show that CADS is insufficient for preventing replication. In contrast, ProCreate significantly lowers the perceptual similarity and Top-1 SSCD scores in all examples, showing that their generated samples are sufficiently different in SSCD score (< 0.4) from all other images in LAION12M. This can be explained by ProCreate's ability to dynamically select the closest LAION12M example to be propelled from during guidance so that the generated image would always be guided away from an arbitrary LAION12M image that it is close to. Since DreamSim imitates human perception for detecting similarities in images, using it as our similarity embedding network ensures that ProCreate outputs do not replicate training images.

Top-1 SSCD Scores. In Figure 8, we compare the percentage of top1-SSCD scores of each 10k images generated with DDIM, CADS, and ProCreate sampling. ProCreate reduces the percentages of DDIM generations by more than half, demonstrating superior ability in preventing data replication from the pre-trained Stable Diffusion model.

Comparison on Distribution Metrics. Interestingly, Table 3 shows that ProCreate also reduced both FID and KID when compared to the baseline DDIM



Fig. 9: Failure cases from the LAION12M experiment in Section 6.

sampling while CADS does not, indicating that ProCreate not only reduces data replication but also improves the generation quality with higher sample diversity.

7 Conclusion and Discussion

We propose ProCreate, a simple method for improving the generation diversity and creativity of diffusion models given a set of reference images. We focus on two setups to demonstrate the effectiveness of ProCreate. First, in the low-data fine-tuning regime, we introduce a new text-to-image few-shot generation dataset FSCG-8 and we show that ProCreate achieves the best performance in terms of sample diversity and distribution metrics compared to prior approaches. Second, for pre-trained diffusion networks, we show that ProCreate can effectively mitigate training data replication and improve sample quality on the LAION dataset. In the future, we expect that ProCreate can be applied to more modalities of data, including audio and video, to promote more diverse and creative generation of digital content.

Broader Impact. This research presents significant broader impacts. It offers content creators and designers the tools to enhance AI-assisted design with a smaller risk of replicating reference images, or private/copyrighted training images. Although the primary implications are beneficial, there exists a potential for this technology to facilitate the design of counterfeit products. Addressing the ethical use and regulatory oversight of such advancements warrants further discussion in future works.

Limitations. Currently, there are several limitations to ProCreate. The guidance process is slow compared to direct generation since it performs Multi-Step Look Ahead and needs to backpropagate through a similarity network. In Table 4, we show ProCreate's generation time at different n_{step} values with batch size 1 and 50 inference steps on an NVIDIA A100 GPU. The method can also sometimes be sensitive to the guidance strength parameter γ, requiring extra hyperparameter tuning on new datasets. Figure 9 shows two failure scenarios where ProCreate sample replicates the Top-1 SSCD matched training image in the first row due to low guidance strength, and where ProCreate's sample quality degrades when the guidance strength is too high in the second row.

Acknowledgement

We thank Zhun Deng and members of the NYU Agentic Learning AI Lab for their helpful discussions. The NYU High-Performance Computing resources, services, and staff expertise supported the compute.

References

1. Bai, A., Hsieh, C., Kan, W.C., Lin, H.: Reducing training sample memorization in gans by training with memorization rejection. CoRR **abs/2210.12231** (2022)
2. Bai, C., Lin, H., Raffel, C., Kan, W.C.: On training sample memorization: Lessons from benchmarking generative modeling with a large-scale competition. CoRR **abs/2106.03062** (2021)
3. Bansal, A., Chu, H., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: CVPR (2023)
4. Benigmim, Y., Roy, S., Essid, S., Kalogeiton, V., Lathuilière, S.: One-shot unsupervised domain adaptation with personalized diffusion models. In: CVPR (2023)
5. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD gans. In: ICLR (2018)
6. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. In: USENIX Security Symposium (2023)
7. Corso, G., Xu, Y., Bortoli, V.D., Barzilay, R., Jaakkola, T.S.: Particle guidance: non-i.i.d. diverse sampling with diffusion models. CoRR **abs/2310.13102** (2023)
8. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
9. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2024)
10. Du, Y., Durkan, C., Strudel, R., Tenenbaum, J.B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., Grathwohl, W.S.: Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In: ICML (2023)
11. Farshad, A., Yeganeh, Y., Chi, Y., Shen, C., Ommer, B., Navab, N.: Scenegenie: Scene graph guided diffusion models for image synthesis. In: ICCV - Workshops (2023)
12. Friedman, D., Dieng, A.B.: The vendi score: A diversity evaluation metric for machine learning. CoRR **abs/2210.02410** (2022)
13. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In: NeurIPS (2023)
14. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023)
15. Giannone, G., Nielsen, D., Winther, O.: Few-shot diffusion models (2022)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
17. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Comput. **14**(8), 1771–1800 (2002)
18. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A.A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen video: High definition video generation with diffusion models. CoRR **abs/2210.02303** (2022)

19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
20. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. **23**, 47:1–47:33 (2022)
21. Ho, J., Salimans, T.: Classifier-free diffusion guidance. CoRR **abs/2207.12598** (2022)
22. Ho, J., Salimans, T., Gritsenko, A.A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. CoRR **abs/2204.03458** (2022)
23. Kingma, D.P., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. CoRR **abs/2107.00630** (2021)
24. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009)
25. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023)
26. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models (2019)
27. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. In: ICLR (2022)
28. Lu, H., Tunanyan, H., Wang, K., Navasardyan, S., Wang, Z., Shi, H.: Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In: CVPR (2023)
29. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Gool, L.V.: Repaint: Inpainting using denoising diffusion probabilistic models. CoRR **abs/2201.09865** (2022)
30. Nagarajan, V.: Theoretical insights into memorization in gans (2019)
31. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In: ICML (2022)
32. Pizzi, E., Roy, S.D., Ravindra, S.N., Goyal, P., Douze, M.: A self-supervised descriptor for image copy detection. In: CVPR (2022)
33. Qu, W., Shao, Y., Meng, L., Huang, X., Xiao, L.: A conditional denoising diffusion probabilistic model for point cloud upsampling. CoRR **abs/2312.02719** (2023)
34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. CoRR **abs/2204.06125** (2022)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
36. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
37. Sadat, S., Buhmann, J., Bradley, D., Hilliges, O., Weber, R.M.: CADS: unleashing the diversity of diffusion models through condition-annealed sampling. CoRR **abs/2310.17347** (2023)
38. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models. In: SIGGRAPH (2022)
39. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Lopes, R.G., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
40. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Lopes, R.G., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.:

Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
41. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: an open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
42. Sehwag, V., Hazirbas, C., Gordo, A., Ozgenel, F., Canton-Ferrer, C.: Generating high fidelity data from low-density regions using diffusion models. In: CVPR (2022)
43. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion art or digital forgery? investigating data replication in diffusion models. In: CVPR (2023)
44. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Understanding and mitigating copying in diffusion models. In: NeurIPS (2023)
45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
46. Song, Y., Kingma, D.P.: How to train your energy-based models. CoRR **abs/2101.03288** (2021)
47. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
48. Wallace, B., Gokul, A., Ermon, S., Naik, N.: End-to-end diffusion latent optimization improves classifier guidance. In: ICCV (2023)
49. Wang, Z., Zhao, L., Xing, W.: Stylediffusion: Controllable disentangled style transfer via diffusion models. In: ICCV (2023)
50. Yang, L., Huang, Z., Song, Y., Hong, S., Li, G., Zhang, W., Cui, B., Ghanem, B., Yang, M.: Diffusion-based scene graph to image generation with masked contrastive pre-training. CoRR **abs/2211.11138** (2022)
51. Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: LION: latent point diffusion models for 3d shape generation. In: NeurIPS (2022)
52. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: CVPR (2023)
53. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: CVPR (2023)
54. Zhu, J., Ma, H., Chen, J., Yuan, J.: Domainstudio: Fine-tuning diffusion models for domain-driven image generation using limited data. CoRR **abs/2306.14153** (2023)

8 Additional Results

Varying Multi-Step Look Ahead n_{step}. Section 3 shows that at each denoising step, $\hat{x}_0$ can be predicted in a single step by Equation 1. However, the results displayed in Figure 10 indicate that the 1-step prediction may be blurry and inaccurate. As more DDIM sampling steps are taken, the image quality increases, leading to an improvement in the similarity embedding and the propulsive guidance gradient. Table 5 further illustrates that while ProCreate still outperforms DDIM sampling when the Multi-Step Look Ahead method is not used ($n_{step} = 1$), the method significantly improves FID, KID, and diversity-focused metrics at higher n_{step} settings.



Fig. 10: Predictions for "Burberry Tote Bag" using different numbers of DDIM sampling steps from the evaluated Burberry checkpoint in Table 1.

Table 5: The effect of Multi-Step Look Ahead n_{step} on ProCreate's Performance. $n_{step} = 0$ represents baseline DDIM sampling that does not use ProCreate.

MSLA n_{step}	FID ↓	KID ↓	Precision ↑	Recall ↑	MSS ↓	Vendi ↑	Prompt Fid. ↑
0	35.11 ± 2.60	4.06 ± 0.48	**0.69 ± 0.06**	0.71 ± 0.05	0.18 ± 0.00	26.15 ± 0.56	**0.34 ± 0.00**
1	29.99 ± 1.49	2.64 ± 0.08	0.60 ± 0.06	0.91 ± 0.05	0.13 ± 0.01	30.90 ± 0.69	0.33 ± 0.00
3	17.82 ± 2.99	0.79 ± 0.47	0.64 ± 0.02	0.91 ± 0.04	0.10 ± 0.01	33.51 ± 0.58	0.33 ± 0.00
5	**14.10 ± 1.64**	**0.72 ± 0.11**	0.66 ± 0.06	**0.97 ± 0.02**	**0.09 ± 0.00**	**34.12 ± 0.36**	0.33 ± 0.00

Using Different Base Diffusion Samplers. We experiment with different base diffusion samplers using the standard fine-tuned "Burberry Designs" checkpoint from Table 1. Table 6 shows that although the DDPM and PNDM base samplers does not perform as well as the DDIM sampler on generative modeling metrics, we can apply ProCreate to them to improve their output diversity while maintaining sample quality.

Varying the Number of Training Samples. We conduct a quantitative evaluation for model checkpoints that are trained on 5 and 25 samples of each FSCG-8 subset respectively. In both cases, the models are trained for 2000 iterations. For the 5-shot fine-tuning runs, there are 45 validation samples, while for the 25-shot fine-tuning runs, there are 25 validation samples. We generate the same number of images as the number of validation samples in each evaluation run. The results are presented in Table 7 and 8. As shown in the tables, ProCreate achieves the best performance in terms of matching the validation distribution (FID, KID) and generating diverse samples (Recall, MSS, Vendi Score). ProCreate also remains competitive in quality-focused metrics (Precision and Prompt Fidelity).

Table 6: Quantitative comparison between DDIM, CADS, and ProCreate on different base samplers DDIM, DDPM, and PNDM. The evaluation is performed with standard fine-tuning, "Burberry Designs" images, and a 10 training and 40 validation split.

Sampler	Method	FID ↓	KID ↓	Precision ↑	Recall ↑	MSS ↓	Vendi ↑	Prompt Fid. ↑
DDIM	Default	35.11 ± 2.60	4.06 ± 0.48	0.69 ± 0.06	0.71 ± 0.05	0.18 ± 0.00	26.15 ± 0.56	**0.34 ± 0.00**
	CADS	34.39 ± 1.97	4.27 ± 0.41	**0.70 ± 0.07**	0.81 ± 0.06	0.18 ± 0.00	25.97 ± 0.20	**0.34 ± 0.00**
	ProCreate	**14.10 ± 1.64**	**0.72 ± 0.11**	0.66 ± 0.06	**0.97 ± 0.02**	**0.10 ± 0.00**	**33.51 ± 0.36**	0.33 ± 0.00
DDPM	Default	37.29 ± 1.97	4.08 ± 0.27	**0.68 ± 0.05**	0.70 ± 0.07	0.19 ± 0.01	26.05 ± 0.68	**0.34 ± 0.00**
	CADS	37.24 ± 1.81	4.36 ± 0.31	**0.68 ± 0.03**	0.73 ± 0.08	0.19 ± 0.01	25.62 ± 0.48	**0.34 ± 0.00**
	ProCreate	**14.80 ± 2.18**	**2.03 ± 0.22**	0.67 ± 0.08	**0.95 ± 0.03**	**0.12 ± 0.01**	**31.74 ± 0.62**	**0.34 ± 0.00**
PNDM	Default	60.86 ± 2.22	8.94 ± 0.53	**0.64 ± 0.09**	0.49 ± 0.10	0.22 ± 0.01	24.51 ± 1.56	**0.30 ± 0.01**
	CADS	63.24 ± 1.47	9.44 ± 0.42	**0.64 ± 0.04**	0.56 ± 0.06	0.23 ± 0.01	23.50 ± 0.54	**0.30 ± 0.00**
	ProCreate	**28.64 ± 2.09**	**2.22 ± 0.34**	0.62 ± 0.05	**0.92 ± 0.04**	**0.11 ± 0.01**	**32.08 ± 0.44**	**0.30 ± 0.00**

DreamBooth Qualitative Samples. In Figure 11, we perform DreamBooth fine-tuning with the experiment setup from Section 5.2 to produce qualitative results with DDIM, CADS, and ProCreate sampling methods. Similarly to the qualitative comparisons in Figure 4, ProCreate generates the most diverse samples that follow the prompts yet do not replicate the Top-1 SSCD matched training images.

Table 7: Quantitative comparison between DDIM, CADS, and ProCreate applied on standard fine-tuning checkpoints for 5-shot learning on various generative modeling metrics. We show the mean and standard deviation values over 5 repeated runs in each cell.

Subset	Method	FID ↓	KID ↓	Precision ↑	Recall ↑	MSS ↓	Vendi ↑	Prompt Fid. ↑
Amedeo	DDIM	11.27 ± 0.69	2.02 ± 0.18	**0.81 ± 0.03**	0.44 ± 0.12	0.36 ± 0.01	14.32 ± 0.70	**0.34 ± 0.00**
	CADS	12.51 ± 0.68	2.43 ± 0.22	0.80 ± 0.04	0.39 ± 0.08	0.36 ± 0.02	14.78 ± 0.96	**0.34 ± 0.00**
	ProCreate	**10.23 ± 0.66**	**0.21 ± 0.03**	0.69 ± 0.06	**0.55 ± 0.17**	**0.19 ± 0.01**	**30.46 ± 0.84**	0.33 ± 0.01
Apple	DDIM	19.68 ± 2.39	0.36 ± 0.13	0.63 ± 0.04	0.70 ± 0.12	**0.11 ± 0.00**	36.11 ± 0.34	**0.30 ± 0.00**
	CADS	27.83 ± 1.34	0.83 ± 0.11	0.48 ± 0.02	0.71 ± 0.12	**0.11 ± 0.01**	36.65 ± 0.75	0.29 ± 0.00
	ProCreate	**14.55 ± 1.49**	**0.12 ± 0.03**	**0.64 ± 0.05**	**0.78 ± 0.10**	**0.11 ± 0.01**	**36.70 ± 0.63**	**0.30 ± 0.00**
Burberry	DDIM	10.43 ± 0.58	0.37 ± 0.02	**0.75 ± 0.04**	0.73 ± 0.05	0.30 ± 0.01	16.80 ± 0.45	**0.35 ± 0.00**
	CADS	12.31 ± 0.69	0.46 ± 0.04	0.69 ± 0.03	0.69 ± 0.04	0.30 ± 0.01	16.70 ± 0.68	**0.35 ± 0.00**
	ProCreate	**9.61 ± 0.51**	**0.27 ± 0.04**	0.69 ± 0.04	**0.89 ± 0.07**	**0.21 ± 0.01**	**23.31 ± 0.56**	**0.35 ± 0.00**
Frank	DDIM	8.76 ± 0.39	1.31 ± 0.11	0.99 ± 0.01	0.52 ± 0.05	0.23 ± 0.00	20.81 ± 0.55	**0.29 ± 0.00**
	CADS	10.06 ± 0.28	1.61 ± 0.04	**1.00 ± 0.01**	0.55 ± 0.07	0.24 ± 0.01	21.18 ± 0.52	**0.29 ± 0.00**
	ProCreate	**5.09 ± 0.33**	**0.57 ± 0.14**	0.93 ± 0.03	**0.62 ± 0.10**	**0.16 ± 0.01**	**29.41 ± 1.66**	0.28 ± 0.00
Nouns	DDIM	2.88 ± 0.43	0.03 ± 0.00	0.12 ± 0.05	0.71 ± 0.17	0.48 ± 0.01	11.90 ± 0.44	**0.24 ± 0.00**
	CADS	3.11 ± 0.46	0.03 ± 0.00	0.11 ± 0.02	0.63 ± 0.12	0.48 ± 0.01	11.78 ± 0.41	**0.24 ± 0.00**
	ProCreate	**2.64 ± 0.31**	**0.02 ± 0.00**	**0.15 ± 0.05**	**0.79 ± 0.11**	**0.44 ± 0.01**	**13.65 ± 0.45**	**0.24 ± 0.00**
Onepiece	DDIM	9.17 ± 0.21	1.24 ± 0.03	0.72 ± 0.06	0.20 ± 0.03	0.29 ± 0.01	22.56 ± 0.41	**0.30 ± 0.00**
	CADS	10.56 ± 0.99	1.52 ± 0.17	0.70 ± 0.05	0.19 ± 0.04	0.30 ± 0.01	22.46 ± 0.51	**0.30 ± 0.00**
	ProCreate	**5.57 ± 0.64**	**0.58 ± 0.10**	**0.73 ± 0.06**	**0.42 ± 0.06**	**0.25 ± 0.01**	**26.67 ± 0.62**	**0.30 ± 0.00**
Pokemon	DDIM	**13.23 ± 0.43**	0.24 ± 0.02	0.46 ± 0.08	0.84 ± 0.07	0.28 ± 0.01	21.94 ± 0.58	**0.31 ± 0.00**
	CADS	17.05 ± 0.40	0.29 ± 0.04	0.44 ± 0.03	0.76 ± 0.12	0.29 ± 0.01	21.98 ± 0.76	**0.31 ± 0.00**
	ProCreate	13.54 ± 0.63	**0.22 ± 0.01**	**0.47 ± 0.09**	**0.88 ± 0.07**	**0.23 ± 0.00**	**23.13 ± 0.22**	**0.31 ± 0.00**
Rococo	DDIM	29.56 ± 1.66	7.01 ± 0.51	0.93 ± 0.03	0.45 ± 0.07	0.20 ± 0.00	19.09 ± 0.32	0.30 ± 0.00
	CADS	33.52 ± 0.63	8.13 ± 0.18	**0.94 ± 0.02**	0.52 ± 0.09	0.21 ± 0.00	18.74 ± 0.57	0.30 ± 0.00
	ProCreate	**23.37 ± 0.36**	**5.07 ± 0.24**	**0.94 ± 0.03**	**0.80 ± 0.04**	**0.17 ± 0.01**	**33.74 ± 0.82**	**0.31 ± 0.00**

Table 8: Quantitative comparison between DDIM, CADS, and ProCreate applied on standard fine-tuning checkpoints for 25-shot learning on various generative modeling metrics. We show the mean and standard deviation values over 5 repeated runs in each cell.

Subset	Method	FID ↓	KID ↓	Precision ↑	Recall ↑	MSS ↓	Vendi ↑	Prompt Fid. ↑
Amedeo	DDIM	14.87 ± 1.42	3.01 ± 0.37	**0.66 ± 0.07**	0.69 ± 0.19	0.30 ± 0.01	14.07 ± 0.41	**0.34 ± 0.00**
	CADS	16.51 ± 0.98	3.48 ± 0.32	0.58 ± 0.08	0.57 ± 0.19	0.32 ± 0.01	13.68 ± 0.26	**0.34 ± 0.00**
	ProCreate	**12.22 ± 1.04**	**2.89 ± 0.24**	0.41 ± 0.07	**0.75 ± 0.05**	**0.29 ± 0.01**	**16.03 ± 0.30**	0.33 ± 0.01
Apple	DDIM	45.35 ± 2.07	2.77 ± 0.22	**0.58 ± 0.03**	0.76 ± 0.08	0.18 ± 0.01	17.85 ± 0.54	**0.31 ± 0.00**
	CADS	30.89 ± 0.35	2.73 ± 0.09	**0.58 ± 0.08**	0.78 ± 0.05	0.19 ± 0.01	17.13 ± 0.52	**0.31 ± 0.00**
	ProCreate	**25.16 ± 1.07**	**2.48 ± 0.14**	**0.58 ± 0.05**	**0.80 ± 0.08**	**0.14 ± 0.01**	**20.01 ± 0.64**	**0.31 ± 0.00**
Burberry	DDIM	16.44 ± 0.18	1.39 ± 0.19	**0.95 ± 0.02**	0.82 ± 0.12	0.27 ± 0.01	13.76 ± 0.42	**0.34 ± 0.00**
	CADS	17.87 ± 1.91	1.51 ± 0.25	0.94 ± 0.03	0.73 ± 0.09	0.26 ± 0.01	14.06 ± 0.42	**0.34 ± 0.00**
	ProCreate	**6.72 ± 1.84**	**0.82 ± 0.26**	0.89 ± 0.07	**0.93 ± 0.07**	**0.16 ± 0.01**	**20.38 ± 0.61**	**0.34 ± 0.00**
Frank	DDIM	6.13 ± 0.28	0.62 ± 0.09	**0.98 ± 0.00**	0.81 ± 0.04	**0.14 ± 0.00**	19.77 ± 0.47	**0.31 ± 0.00**
	CADS	**4.57 ± 0.10**	0.60 ± 0.07	0.96 ± 0.00	0.82 ± 0.04	**0.14 ± 0.00**	19.50 ± 0.21	**0.31 ± 0.00**
	ProCreate	4.90 ± 0.19	**0.34 ± 0.02**	0.93 ± 0.09	**0.83 ± 0.03**	**0.14 ± 0.00**	**21.73 ± 0.26**	**0.31 ± 0.01**
Nouns	DDIM	1.81 ± 0.09	0.02 ± 0.00	0.59 ± 0.09	0.95 ± 0.05	0.50 ± 0.01	8.58 ± 0.14	**0.25 ± 0.00**
	CADS	1.24 ± 0.05	0.02 ± 0.00	**0.66 ± 0.12**	0.97 ± 0.05	0.51 ± 0.01	8.34 ± 0.18	**0.25 ± 0.00**
	ProCreate	**1.22 ± 0.01**	**0.01 ± 0.00**	0.62 ± 0.08	**0.98 ± 0.03**	**0.46 ± 0.01**	**14.55 ± 0.19**	**0.25 ± 0.00**
Onepiece	DDIM	2.85 ± 0.13	0.10 ± 0.02	**0.83 ± 0.04**	0.69 ± 0.05	**0.25 ± 0.00**	16.86 ± 0.15	0.30 ± 0.00
	CADS	3.02 ± 0.16	0.15 ± 0.01	0.78 ± 0.03	0.75 ± 0.02	0.26 ± 0.01	16.81 ± 0.29	**0.31 ± 0.00**
	ProCreate	**2.14 ± 0.11**	**0.09 ± 0.01**	0.82 ± 0.04	**0.79 ± 0.01**	**0.25 ± 0.01**	**18.05 ± 0.45**	0.30 ± 0.00
Pokemon	DDIM	6.32 ± 0.34	**0.05 ± 0.01**	0.78 ± 0.05	0.78 ± 0.05	0.31 ± 0.01	14.89 ± 0.44	**0.31 ± 0.00**
	CADS	8.20 ± 0.30	0.06 ± 0.00	**0.86 ± 0.03**	0.74 ± 0.03	0.31 ± 0.02	14.77 ± 0.67	**0.31 ± 0.00**
	ProCreate	**4.88 ± 0.93**	**0.05 ± 0.00**	0.82 ± 0.05	**0.86 ± 0.03**	**0.28 ± 0.01**	**15.95 ± 0.42**	**0.31 ± 0.00**
Rococo	DDIM	10.30 ± 0.58	1.82 ± 0.09	0.83 ± 0.02	0.94 ± 0.02	0.15 ± 0.00	15.08 ± 0.37	**0.34 ± 0.00**
	CADS	13.36 ± 0.61	1.85 ± 0.11	0.83 ± 0.02	0.92 ± 0.00	0.15 ± 0.00	15.15 ± 0.34	**0.34 ± 0.00**
	ProCreate	**9.41 ± 1.06**	**1.63 ± 0.07**	**0.86 ± 0.03**	**0.97 ± 0.01**	**0.14 ± 0.01**	**15.89 ± 0.84**	**0.34 ± 0.00**



Fig. 11: Qualitative comparison between DDIM, CADS, and ProCreate for few-shot creative generation on FSCG-8 with DreamBooth fine-tuning. For each sampling method, we show two prompts and four generated samples for each prompt. In addition, we match each sample from ProCreate with its closest training image based on the SSCD score [32] between the matched pair.