Specialist Diffusion: Plug-and-Play Sample-Efficient Fine-Tuning of Text-to-Image Diffusion Models to Learn Any Unseen Style

Haoming Lu[1*], Hazarapet Tunanyan[1*], Kai Wang[2], Shant Navasardyan[1], Zhangyang Wang[1,3], Humphrey Shi[1,2]
[1]Picsart AI Research (PAIR), [2]U of Oregon, [3]UT Austin

https://github.com/Picsart-AI-Research/Specialist-Diffusion

Abstract

Diffusion models have demonstrated impressive capability of text-conditioned image synthesis, and broader application horizons are emerging by ***personalizing*** *those pre-trained diffusion models toward generating some specialized target object or style. In this paper, we aim to learn an unseen style by simply fine-tuning a pre-trained diffusion model with a handful of images (e.g., less than 10), so that the fine-tuned model can generate high-quality images of arbitrary objects in this style. Such extremely low-shot fine-tuning is accomplished by a novel toolkit of fine-tuning techniques, including text-to-image customized data augmentations, a content loss to facilitate content-style disentanglement, and sparse updating that focuses on only a few time steps. Our framework, dubbed* ***Specialist Diffusion****, is* ***plug-and-play*** *to existing diffusion model backbones and other personalization techniques. We demonstrate it to outperform the latest few-shot personalization alternatives of diffusion models such as Textual Inversion [7] and DreamBooth [24], in terms of learning highly sophisticated styles with ultra-sample-efficient tuning. We further show that Specialist Diffusion can be integrated on top of textual inversion to boost performance further, even on highly unusual styles. Our codes are available at:* https://github.com/Picsart-AI-Research/Specialist-Diffusion.

1. Introduction

Image synthesis has received increasing attention, partially owing to the recent breakthroughs made by diffusion models [10, 23, 28, 30, 34]. Training a diffusion model requires gradually adding random noises with a sequence of diffusion steps, and learning to rebuild data from noises by reversing the steps. By running the diffusion process on a lower-dimensional latent space instead of the pixel space, the latent diffusion model [23] achieves competitive performance more efficiently. Diffusion models are capable of both unconditioned and conditioned image synthesis, the previous one generates samples from random noise (noise-to-image), while the latter takes a condition such as text to guide the generation. In this paper, we will focus on text-to-image synthesis for its broad application interests.

Diffusion models nowadays are able to synthesize high-quality images on a wide variety of sophisticated text

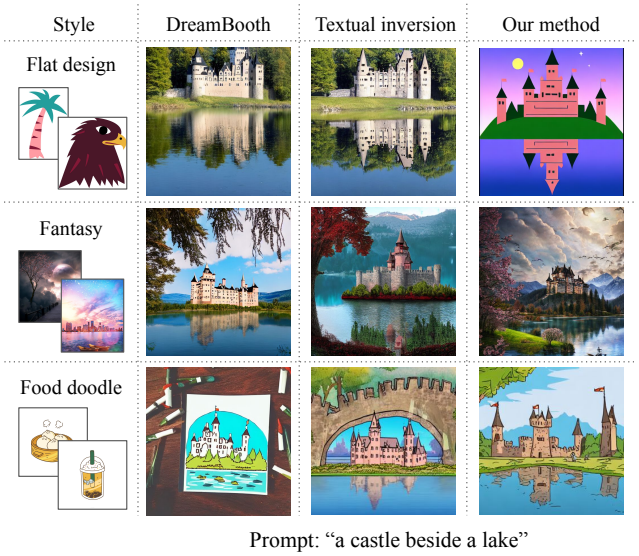

Figure 1. Comparison of fine-tuning the pre-trained Stable Diffusion [23] model, using our Specialist Diffusion method versus two other methods, from left to right columns. Three rows represent three different, rare styles (*"Flat design"*, *"Fantasy"*, and *"Food doodle"*) that we hope to personalize the Stable Diffusion model to learn, using only a handful of samples (even less than 10). All examples are generated using the same text prompt except for the style identifier. As a method focused on representing new objects, DreamBooth [24] performs poorly when being applied to capturing styles. Textual inversion [7] achieves neat performance on some styles, but fails on more unusual styles such as "Flat design". Specialist Diffusion (rightmost) succeeds to capture those highly unusual, specialized, and sophisticated styles via few-shot tuning. Please see Sec. 4 for more dataset and experiment details.

[1]The first two authors Lu and Tunanyan contributed equally.

prompts. However, those models still have limited coverages on what they can generate - and their generated images are often found to miss highly specialized objects or unusual styles [19]. Practitioners are hence motivated to **personalize** those models on the target data of their interest. Notably, many target domains have fine granularity, such as selfies of one user, or a specialized design style; consequently, samples available from such target domains would be limited and constitute **few-shot** scenarios.

There was little work on few-shot personalization with diffusion models until recently. Most conditional models focus on specific applications. Here we consider the general case of few-shot generation for unknown classes at test time. One naive solution is to fine-tune the model, but it can easily go sample-costly due to the enormous model size. For example, the training of Waifu-diffusion [15], a model fine-tuned from Stable Diffusion v1.4 [23], took approximately 10 days on 8 Nvidia A40 GPUs with 680K text-image pairs. Some initial attempts were made to reduce the sample complexity. D2C [26] shows that latent space and self-supervised learning lead to few-shot tuning with as few as 100 samples. [8] explicitly conditions the denoising diffusion dynamics on a support set of target domain samples, and takes only 5 samples to generate images from a new class. Despite the promise, they modified the model architecture by injecting an unconventional set-based vision transformer (ViT) [5] to aggregate new image patch information. Moreover, their evaluation was on unconditional image generation with the resolution of 32×32 or 64×64, and it remains unclear how their method can extend to text-to-image synthesis, or to generating high-resolution images.

Several latest works shed new light on the horizon of few-shot personalization. Among them, DreamBooth [24] fine-tunes the pre-trained diffusion model to bind a unique identifier with an unseen object. Textual Inversion [7], in comparison, learns to represent a new concept through a new token "word" in the embedding space, without fine-tuning the parameters of the pretrained model. Both methods can achieve personalization with $\sim$ 5 images. However, both DreamBooth and Textual inversion were mainly demonstrated to synthesize new **unknown objects** (such as a selfie) in various *known styles* or contexts; while their extension to the complementary side, e.g., synthesizing *known objects* in new **unknown styles**, are under-explored and often found to be unsatisfactory, despite the apparent demands of capturing unfamilar or personalized styles by artist users.

In this paper, we focus on *fine-tuning* a pre-trained text-to-image diffusion model, to learn a highly specialized *unseen style*, using a handful of images. Our goal is to have the fine-tuned model generate high-quality images of arbitrary (known) objects in this (previously unknown) style. Our proposed solution, **Specialist Diffusion**, is a plug-and-play set of fine-tuning techniques that works with any diffusion backbone without altering their architectures, and can also be integrated with existing personalization methods such as [7, 24]. Firstly, our customized augmentations are specifically designed for the text-to-image scenario, that augment not only the images but also the text prompts with prior language knowledge from the image augmentation. Secondly, to prevent the model from overfitting to the target style while losing generalization to various objects, we introduce a content loss to preserve the ability to generate specified content based on CLIP [21]. Besides, instead of updating all diffusion steps, we find that updating only a sparse subset of steps can significantly improve the few-shot training efficiency, with comparable or sometimes better fine-tuning performance. Our contributions are summarized as:

- Specialist Diffusion is a general, plug-and-play framework to fine-tune pre-trained diffusion models with few-shot samples (less than 10), to learn highly sophisticated and unusual styles, without bells and whistles.
- We propose a rich set of techniques including various customized data augmentations that augment both text prompts and images, a content loss for disentanglement, and sparse diffusion step updating for both sample and computation efficiency.
- Specialist Diffusion not only improves upon the state-of-the-art diffusion model personalization methods such as Textual Inversion [7] and DreamBooth [24] (e.g., see Fig. 1), but also can be combined on top of them to jointly boost style personalization further.

2. Related Work

2.1. Diffusion Models

Given a sample x_0 from an unknown $q(x_0)$ distribution, the goal of diffusion models [10, 28, 29] is to learn a parametric model $p_\theta(x_0)$ to approximate the original $q(x_0)$ distribution. The model $q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1})$ gradually adds noise to the input and is called *forward process*. During the backward generative process, the model will obtain a clean version of x_t by removing the noise from the x_{t+1} timestep. To approximate $q(x_0)$ with the $p_\theta(x_0)$, θ can be learned by optimizing the following objective

$$L(\theta) = \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,I), t} \left[||\epsilon_\theta(x_t, t) - \epsilon||^2 \right] \quad (1)$$

Where $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon_t$ and $\epsilon_\theta(x_t, t)$ is a trainable U-Net like architecture to approximate the noise.

As the objective in Eq. (1) does not depend on any specific connections between $x_1, ..., x_T$ latent variables, one can construct a new model $p_\theta(x_0)$ such that the $q(x_0)$ can be modeled with p_θ without additional training. This key idea of DDIM [29] offers a new way of deterministic diffusion process that also speeds up sampling during inference.

Training powerful DMs for high-resolution image synthesis takes too many GPU days. To lower the computational demand and stablize the generalization towards high resolution, Rombach *et al.* [23] applied diffusion denoising operation on the latent space of a pre-trained encoder. They used the encoder of VQGAN [6] to compress high-resolution images into a compact latent space, and the decoder to reconstruct them from the denoised latent variable. Given an image x_0 in RGB space, the encoder $\mathcal{E}$ encodes the x_0 into a latent representation $z = \mathcal{E}(x_0)$. The objective then only considers the latent variable z instead of x:

$$L_{ldm}(\theta) = \mathbb{E}_{z\sim\mathcal{E}(x_0),\epsilon\sim\mathcal{N}(0,I),t}\left[||\epsilon_\theta(z_t, t) - \epsilon||^2\right] \quad (2)$$

Operating DMs on smaller latent spaces significantly decreased the processing time for training and inference, and delivers current state-of-the-art (SOTA) results on high-resolution image synthesis (under the name of Stable Diffusion). Therefore, throughout this paper we choose to benchmark all few-shot personalization methods with the latest Stable Diffusion model [23], although they can be plugged in other diffusion model backbones the same way.,

2.2. Few-Shot Generation

The majority of large generative models are trained on millions or even billions of examples, which is often not feasible to acquire for many fine-grained or proprietary image domains. Few-shot image generation [13, 16, 18, 33] targets solving this issue by using only a handful of examples to generate images. Past works in VAEs or GANs apply fine-tuning to a pre-trained GAN generator, selecting coresets of training data, leveraging differentiable or adaptive data augmentations, or applying fine-tuning weight regularizers [1, 3, 11, 13, 17, 22, 27, 31, 35, 36].

Few-shot diffusion models are inspired by similar approaches. Sinha *et al.* [26] apply self-supervised learning with diffusion process to learn an unconditional model for few-shot conditional image synthesis. To compensate for the absence of conditioning during training, they train an additional model over the latent representation of the inputs, which will play a guiding role during inference. Liu *et al.* [14] designed a unified framework for image synthesis based on reference examples and their method can be applied into unconditional diffusion models. By explicitly conditioning the DDPM module, Giannone *et al.* [8] used a set-based Vision Transformer (ViT) [5] to extract general information by patches from few-shot examples. lso, they did not consider text-guided image synthesis. Blattmann *et al.* [2] explored solving few-shot diffusion problem by using the image-retrieval augmentation method. For each particular query input x_q, they retrieve a close set of examples from an external dataset by using *top-k* operation over cosine similarity scores in embedding space. This subset is used with an additional conditioning branch.

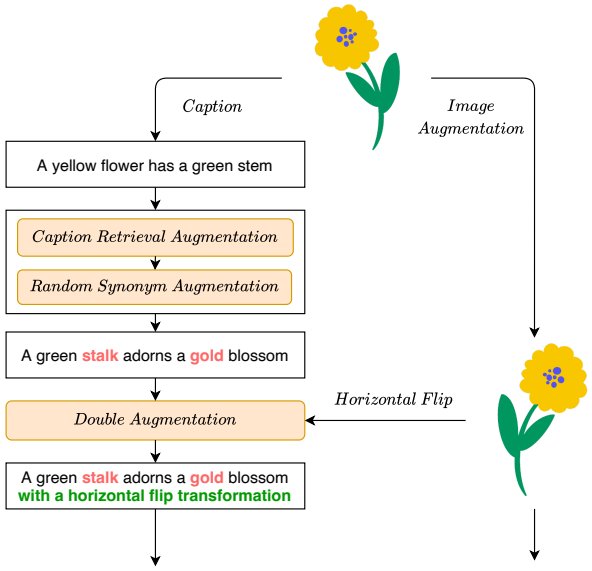

Figure 2. Illustration of our customized data augmentation flow.

Two recent works made particularly notable progress on text-to-image few-shot personalization at high resolution, both built on Stable Diffusion or equivalent. DreamBooth [24] designed a personalized fine-tuning procedure by structuring the prompts with a special form *"a [V] [class noun]"*, where *"[V]"* is a unique identifier of the subject and *"[class noun]"* is the true class of it. They mainly focus on objects in their experiments, and our experiments also observed that DreamBooth lacks effectiveness to capture few-shot styles when applied out of the box. Textual Inversion [7] also learns to generate personalized concepts, including objects or artistic styles, by describing them using new "words" in the embedding space of pre-trained text-to-image models. These can then be used in new sentences, just like any other word. Textual Inversion makes our closest competitor in few-shot style personalization; despite its effectiveness, later in experiments we will demonstrate it falls short of capturing some other very unusual styles, such as "Flat Design" in Fig. 1, from few-shot demonstrations. Also, since Textual Inversion does not fine-tune the pre-trained model weight, it ispotentially complementary to fine-tuning based options such as DreamBooth or ours.

3. Specialist Diffusion: Our Technical Toolkit

Our framework carries a novel toolkit of fine-tuning techniques, including text-to-image customized data augmentations, a content loss to facilitate content-style disentanglement, and sparsely updating diffusion time steps.

3.1. Data Augmentations for Text2Image Diffusion

We separate the augmentation process into two parts: *augmentation of images* and *augmentation of captions*.

Image Augmentation - Image augmentation continues

to be an integral part of solutions to many computer vision problems. Yet when applied to generative problems such as image synthesis, the majority of the functions of geometric transformations or other heavy modifications of inputs lead to a known problem called "augmentation leakage" [11,36]. The generative model often memorizes the training examples and their augmented versions and generates similar images during inference. For example, while many rotated images could be legitimated natural photos, their occurrence in natural image collections has lower probabilities. Hence applying too many rotations in training is observed to bias the chance of generating more rotated objects which are untended. To minimize the augmentation leakage, we skip heavier image augmentations such as AugMix [9] and RandAug [4], and select a set of mild augmentation functions that do not aggressively distort the input image such as randomly flipping or AutoContrast. A full list of our image augmentations can be found in the Supplementary.

Figure 3. An example to show the risk of losing semantic object knowledge when personalizing to a new style. Images are generated with text prompt "two people sitting on a bench in the park", and the target style is "Flat design". Left: original image; middle: tuned for 100 epochs; right: tuned for 1000 epochs.

Text Prompt Augmentation - Many known augmentation methods are designed for images only and the text conditioning part remains untouched, leaving a unique open opportunity. Instead of having one description of an image as a caption, we give different interpretations of the same caption by replacing their words with their synonyms. An extension of this approach is to replace the whole sentence with another retrieved one, which is semantically similar to the original caption. We also modify each caption to correspond to the randomly chosen image augmentation method by extending the text with a description of the selected function. Our multiple levels of caption augmentations, *caption-retrieval augmentation*, *synonym augmentation*, and *doubled augmentation*, are illustrated in Fig. 2 and introduced next. The ablation comparison can be found in Sec. 4.5.

Caption Retrieval Augmentation - First, we replace the whole caption of an image with another description, without changing its semantic meaning. To find similar descriptions for each caption, we compare the caption of an input image, and the texts from the external LAION [25] text set (72 million sentences), using their cosine similarity score in the embedding space of the CLIP [21] text encoder. For all similar external sentences passing cosine similarity threshold (we used an ad-hoc threshold of 0.7 for all), we randomly choose from them as caption augmentations.

Synonym Augmentation - Another effective method to augment text is to randomly replace a word with one of its synonyms. For text-guided image synthesis, it helps the model to interpret an image in different yet equivalent ways. This approach diversifies the representations of captions and generalizes well for a few-shot diffusion scenario. It does not increase the training time visibly, because we use a predefined synonym correspondence for each word we have in the few-shot examples.

Doubled Augmentation - When images are augmented leaving their captions unchanged, the diffusion model relies on images more than their descriptions, which means captions become less informative. We believe that each image modification should follow its caption modification as well. This technique will help reduce augmentation leakage. To correspond to proper augmentation for the captions, we extend them by adding a small description of a randomly chosen augmentation. Those descriptions push the text representations to be discriminative in the embedding space and better attached to those augmentation attributes.

3.2. Content Loss

As the training continues, it is observed that the model gradually loses the semantic knowledge to generate outputs that meet the input text condition such as objects. For example in Fig. 3, when overly tuned, the image start producing wrong semantics such as confused human bodies and ungrounded trees, and missing the park context. To overcome this issue, an additional content loss is introduced to disentangle the knowledge of content (inherited from pre-trained model) and style (learned from few-shot examples), and help preserve the model ability to understand the semantics in condition while learning a new style.

As Specialist Diffusion attempts to learn an abstract style that is applicable to arbitrary content afterward, we want to ensure that the images generated with a text prompt t will fit the content described in t. As our inspiration, StyleCLIP [20] proposed a CLIP-based approach on text-driven image generation. With an image generation network G, the optimal latent w w.r.t the text prompt t is defined with

$$\arg\min_{w \in w^+} D_{CLIP}(G(w), t) + R \tag{3}$$

where D_{CLIP} is the cosine distance between the CLIP [21] embeddings of the image and the text, and R denotes other regularization terms (omitted). While this approach neatly exploits the ability of CLIP to link images to the text semantics, precise text prompts of the training samples are not available in our task. As described in Sec. 3.1 and Sec. 4.1, our text prompts will be generated by image captioning and decorated during the augmentation. In specific,

given a training text-image pair (t^*, I), an image I^* will be synthesized by the current model with the text prompt t^*. Then, instead of directly calculating $D_{CLIP}(I^*, t^*)$, the embedding of the input image I will be used:

$$L_{content}(\theta) = D_{CLIP}(I^*, I) \tag{4}$$

The final loss function will be the weighted sum of LDM loss (Eq. 2) and content loss (λ_c is a hyperparameter):

$$L(\theta) = L_{ldm}(\theta) + \lambda_c \cdot L_{content} \tag{5}$$

We used $\lambda_c = 0.001$ in all our experiments.

3.3. Sparse Updating

In diffusion models, each processing path (forward and backward) takes T timesteps to gradually add or remove noise. To get high-quality results, one often takes a sufficiently larger T, such as $T = 1000$. However, large T causes expensive training and inference. In our few-shot case, we conjecture that not all those steps are unnecessary.

Instead of updating all timesteps during training, we propose to update only a small part of them: 10% or even 1%. We perform sampling from significantly smaller sets, such as $S_1 = \{10, 20, ..., 1000\}$ or $S_2 = \{100, 200, ..., 1000\}$. That will significantly accelerate training and convergence in the first place. Furthermore, we find that by sparse updating, no performance degradation was observed; and sometimes, the visual quality even improves, e.g., preserving more object or background details (see Fig. 8), presumably owing to the regularization effect of sparsity [3].

4. Experiment

4.1. Data Preparation

We collect three few-shot style datasets, that are motivated by real-world applications. Specifically, each dataset is created or collected by our in-house graph designers and represents a unique artwork style that is popular for visual content creation and re-mixing among social media users, hence "personalizing" those styles to users' own artwork makes appealing demands. The examples in each dataset are hand-curated by our expert artists, since massive production of high-quality artwork examples in those categories is infeasible practically, constituting the need of "few-shot". Fig. 4 shows a few examples from them.

- **Flat design** (25 images): 2D flat icon style that generally omits shadows and textures.
- **Fantasy** (15 images): exaggerated color and contrast with fictional feels, that are popular for rendering medieval paintings, digital or sci-fi illustrations.
- **Food doodle** (9 images): hand-drawn cartoon style with certain amounts of shadows and details, which is the most popular for plotting food and drinks.

Figure 4. Samples from three meticulously selected datasets: "Flat design", "Fantasy", and "Food doodle".

For all images, we use BLIP [12] to generate text captions and manually inspect their quality with artists. For example, the first example from the dataset "Fantasy" in Fig. 4 is captioned *"boats floating in front of a cityscape"*.

4.2. Implementation

All the models mentioned in this paper are fine-tuned from Stable Diffusion [23] v1.4 implemented by Diffusers [32]. The method has been tested at both 256×256 and 512×512 resolution. If not specifically stated, the training steps are randomly sampled from a 10% uniform sparse subset of the original 1000 diffusion steps with DDPM [10] scheduler. A default learning rate of $1.0e-5$ gives good results on all datasets tested. Only the parameters of the U-Net will be updated, except the embedding of the additional "word" when combined with Textual inversion [7]. Our method is trained efficiently on a single Nvidia RTX A6000 GPU, with 256×256 and 512×512 input resolutions. We did not add our new augmentations to Dreambooth/Textual-Inversion since those belong to our holistic innovations, and we revealed their incremental gains in ablation study. Baseline hyperparameters are tuned to our best effort.

4.3. Main Results

A prompt for a classic latent diffusion model can be very complex and specific. For example, here is a sample text prompt for Stable Diffusion [23]: *ultrarealistic, (native american old woman) portrait, cinematic lighting, award-winning photo, no color, 80mm lense*. This prompt contains information about lightning, colors, and camera lenses. However, just as the above prompt does not apply to the style specified by the dataset "Flat design", and we observe that arbitrarily adding descriptions about details could be invalid or even conflict with the desired style. A valid

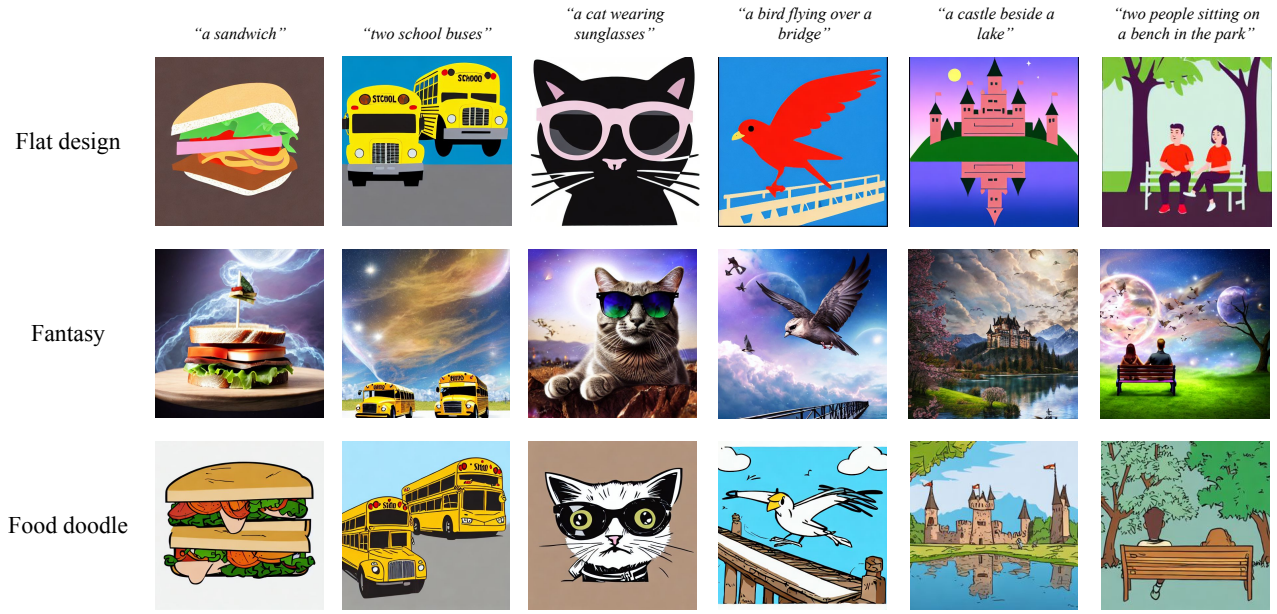

Figure 5. Samples generated by models fine-tuned on the three datasets. The left column shows the dataset on which the model is trained on, and the top row shows the text prompt used to generate the image.

prompt for a style-specific image generation model is supposed to focus more on the intended content. Therefore, the prompts we use for evaluation will be concerned only with the content rather than stylistic details. More discussions about choosing prompts are in the Supplementary.

Fig. 5 shows a collection of generation results to illustrate how our methods learn different unseen styles with extremely low-shot data. Text prompts at different levels of complexity ranging from a single object to a composited scenario have been used. regardless of whether the desired content is complex or close to the distribution of the training data, models fine-tuned with our method successfully generate results satisfying the text condition in target styles.

We also compared our method with other state-of-the-art few-shot methods: DreamBooth [24] and Textual inversion [7]. Fig. 1 and Fig. 6 show the results obtained by applying different methods on the same datasets. The class token is omitted in the figures for easier understanding. For instance, the actual input for our method is "a castle beside a lake in the style of flat" in the top row of Fig. 1, for DreamBooth and Textual inversion it is "a castle beside a lake in the style of [V]". While adapting DreamBooth to styles does not yield promising results (totally ignoring the style or imprecisely presenting), textual inversion capture some styles to a certain extent, but fails catastrophically on some others such as the castle in the Flat design style (Fig. 1) or the bird in Fantasy style (Fig. 6). In comparison, Specialist Diffusion succeeds to capture all those highly unusual, specialized, and sophisticated styles with a few examples.

Style\Method	DB	TextInv	Ours	Ours+Inv
Fantasy	437.254	452.150	399.351	**352.650**
Flat design	445.089	460.252	421.410	**363.362**
Food doodle	491.302	451.466	441.640	**409.607**

Table 1. FIDs on different styles × different methods

As of the time of this paper, there is not a universally recognized **quantitative** metric for assessing the performance of learning a style using a diffusion model. We prompted to generate a hundred images per dataset and compute the following: (1) **FID** averaged across different styles and all images, in Table 1, although it is important to note that calculating FID on a small number of samples is not always reliable; (2) **style loss** in image stylization that estimates the distance between two images based on styles. We treat every sample in each training set as a style reference. Table 2 lists the style loss, averaged over all pairs between {each generated image in this style, each style reference}. (3) **user study** comparing three different models for all datasets, in Figure 9. For each dataset, four subjects voted for the three methods' generated images, based on their style alignment with their source reference dataset.

Style\Method	DB	TextInv	Ours	Ours+Inv
Fantasy	1.087	0.202	0.202	**0.147**
Flat design	0.839	0.276	0.116	**0.098**
Food doodle	0.492	0.104	0.046	**0.037**

Table 2. Average style loss (VGG-based) between generated images & corresponding training examples. Numbers scaled by 10^2.

Figure 6. Continued comparison of our model and other SOTA methods. Random seed is fixed for generation in this figure and Fig. 1.

Figure 7. Combination of our model and textual inversion. Text prompts used for generation are listed top, styles of the corresponding datasets are listed under, and the methods for training the models are listed left.

4.4. Combining with Textual Inversion

Since Textual inversion does not exploit fine-tuning, it is natural to consider how our method could be combined with it: first fine-tuning the pre-trained model using Specialist Diffusion, then generating the style word token using textual inversion from the tuned model, to be used in the prompt for synthesizing new style images. That essentially allows us to *iteratively optimize the model and the prompt*. We verified this cascaded pipeline experimentally in Fig. 7. By integrating textual inversion with our method, the results capture even richer details without losing the style.

4.5. Ablation Study

Overall Method - The combination of all three contributions delivers our best performance. All proposed methods, such as Prompt Augmentation (*double, synonym, and*

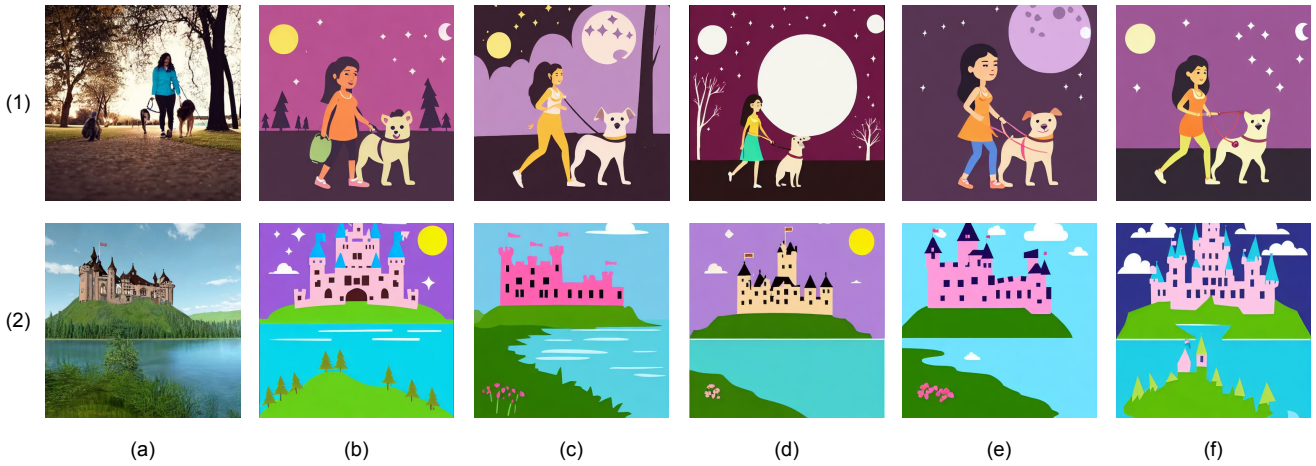

Figure 8. Ablation study of the method. Images are generated with dataset "flat design". Text prompts for each group: (1) "young brown woman walking her dog in a park at night with a full moon"; (2) "a castle beside a lake". Methods for each group: (a) Stable Diffusion; (b) fined-tuned model by our full method; (c) our method w/o caption-retrieval augmentation; (d) our method w/o synonym and caption-retrieval augmentations; (e) our method w/o content loss; (f) our method w/o sparse updating.

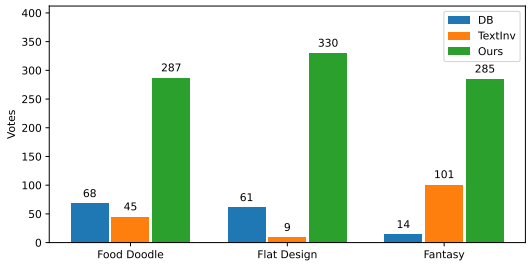

Figure 9. User study with the question: *"which one of the three generated images aligns best with the reference image in style?"*

caption-retrieval), Content Loss, and Sparse Updating complement each other well to provide high-quality few-shot image synthesis. As can be seen from Fig. 8 (b), it contains richer details about the environment, such as trees, stars in the sky, and well-structured bodies of two objects.

Text Prompt Augmentations - Conditioning diversification plays an important role in the generalization. Without those caption augmentations, the model starts to miss many image details and struggles to generalize well for different prompts. For example, Fig. 8 (c) shows that when the caption-retrieval augmentation is dropped, some structures are missed or distorted, such as the moon (upper row) or the peak of the castle (lower). If we further drop both synonym and caption-retrieval augmentations, more objects start to be generated problematically, such as the woman's face and the dog, as well as multiple moons in the sky (upper).

Content Loss - The content loss ensures context preservation during style extraction from few-shot examples. Fig. 8 (e) (upper) supplies good evidence of how the model starts to forget about the main context of the prompt. The word *"park"* is strictly mentioned in the prompt, but the model misses that information during image synthesis.

Sparse Updating - Lastly, updating full timesteps can overfit few-shot examples too. In many of our cases, the sparse updating expresses the style images better. For example, Fig. 8 (f) (upper) contains a visualization of a parkway, however, misses the trees, which are important features of the park. More results at different sparsity levels can be found in this supplementary.

Besides the above visual results, we design a quantitative metric by computing the logits-per-image similarity score between {generated image, text prompt} in the CLIP feature space, averaged across all generated images mentioned in Sec. 4.3. A higher score indicates better-preserving semantics. For results in Fig. 8, such text-image content scores are: **34.7965** (column b), 34.3004 (c), 32.3135 (d), 34.1156 (e), and 33.4526 (f). Those numbers align with our visual impression. Tables 1 and 2 (both last columns) further show the gain from combining Text Inversion and ours.

5. Conclusion

In this work, we presented Specialist Diffusion, a plug-and-play fine-tuning framework to personalize large diffusion models on an unseen style domain with only a small number of images. By introducing customized data augmentation, content loss and sparse updating, Specialist Diffusion reaches both sample and computation efficiency in diffusion model fine-tuning. Experiments have shown that, aside from outperforming SOTA methods, Specialist Diffusion can be combined with them for further improvement. While our work mainly targets artistic and creative editing for benign purposes, it is possible that our method, just like every other generative model, might be abused to generate fake or hateful visual contents. Detecting and rejecting those contents makes important future research.

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. 3

[2] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models. In *Neural Information Processing Systems (NeurIPS), 2022.*, 2022. 3

[3] Tianlong Chen, Yu Cheng, Zhe Gan, Jingjing Liu, and Zhangyang Wang. Data-efficient gan training beyond (just) augmentations: A lottery ticket perspective. *Advances in Neural Information Processing Systems*, 34:20941–20955, 2021. 3, 5

[4] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020. 4

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 3

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021. 3

[7] 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. 1, 2, 3, 5, 6

[8] Giorgio Giannone, Didrik Nielsen, and Ole Winther. Few-shot diffusion models, 2022. 2, 3

[9] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 4

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 5

[11] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 3, 4

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 5

[13] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020. 3

[14] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021. 3

[15] Anthony Mercurio. Waifu diffusion. https://huggingface.co/hakurei/waifu-diffusion, 2022. 2

[16] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[17] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *ICCV*, 2019. 3

[18] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A. Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10743–10752, June 2021. 3

[19] Jonas Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation. *arXiv preprint arXiv:2204.13988*, 2022. 2

[20] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 4

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

[22] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv preprint arXiv:2010.11943*, 2020. 3

[23] 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 (CVPR)*, pages 10684–10695, June 2022. 1, 2, 3, 5

[24] 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. 2022. 1, 2, 3, 6

[25] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 4

[26] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2, 3

[27] Samarth Sinha, Han Zhang, Anirudh Goyal, Yoshua Bengio, Hugo Larochelle, and Augustus Odena. Small-gan: Speeding up gan training using core-sets. In *ICML*, 2020. 3

[28] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1, 2

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[30] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[31] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. *arXiv preprint arXiv:2104.03310*, 2021. 3

[32] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5

[33] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: Effective knowledge transfer from gans to target domains with few images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[34] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022. 1

[35] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained gans for generation with limited data. In *International Conference on Machine Learning*, pages 11340–11351. PMLR, 2020. 3

[36] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020. 3, 4