CUSTOMIZING PRE-TRAINED DIFFUSION MODELS FOR YOUR OWN DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

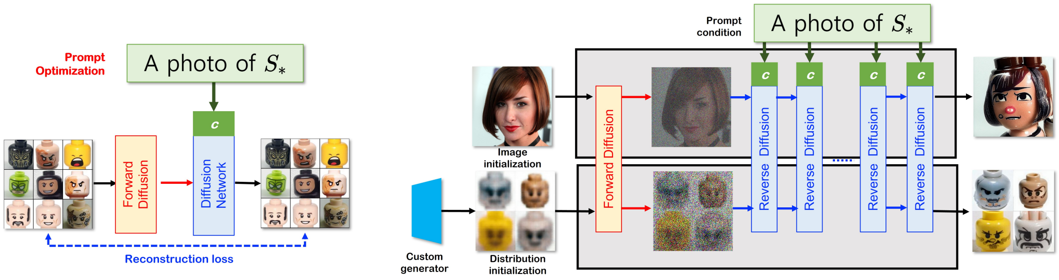
Recently, several large-scale text-to-image diffusion models have been released, showing unprecedented performance. Since the shift from learning a task-specific model from scratch to leveraging pre-trained large-scale models is an inevitable trend in deep generative modeling, it is necessary to develop methods to better utilize these models. In this paper, we propose a method dubbed Diffusion model for Your Own Data (DYOD) that can effectively utilize a pre-trained text-to-image diffusion model to approximate the implicit distribution of a custom dataset. Specifically, we first obtain a text prompt that can best represent the custom dataset through optimization in the semantic latent space of the diffusion model. In order to be able to better control generative image content, in particular geometry of the objects, we show that the text prompt alone is not sufficient, but rather an informative initialization that can guide the pre-trained diffusion model is necessary. As representative examples, we demonstrate that learned distribution initialization from user's data set or an image initialization by user's sketch, photo, etc. serves the goal for customizing diffusion model for user's own data. Experiments show that the customized DYOD outperforms the Stable Diffusion baselines both qualitatively and quantitatively with accelerated sampling speed.

1 INTRODUCTION

Imagine that you want to train a generative model on a custom dataset (maybe you are a designer of a company and want to generate the images of your company's products), but all you have is a single Geforce 1080 Ti. With such limited computational resources, training or fine-tuning a modern generative model is often infeasible. Alternatively, you may want to apply the publicly available pre-trained models to your dataset, but how?

Over the last decade, deep generative modeling has rapidly advanced in image synthesis (Karras et al., 2019; Brock et al., 2018; Child, 2020; Vahdat & Kautz, 2020; Kingma & Dhariwal, 2018). In particular, diffusion models (Ho et al., 2020; Song et al., 2020b; Sohl-Dickstein et al., 2015) have arisen as a powerful class of generative models, achieving remarkable performance in generating images (Dhariwal & Nichol, 2021), videos (Ho et al., 2022), and beyond. Diffusion models are parameter-efficient, trained on a stationary objective function, and scale well. Moreover, thanks to their great modularity, even unconditional diffusion models can serve as generative prior for in-painting, super resolution, colorization, text-guided synthesis, etc (Song et al., 2020b; Chung et al., 2022; Avrahami et al., 2022). Since the authors of GLIDE (Nichol et al., 2021) showed that diffusion models could even scale to hundreds of millions of text and image pairs, several large-scale text-to-image diffusion models have been proposed, showing unprecedented performance (Ramesh et al., 2022; Saharia et al., 2022). Recently, Rombach et al. (2022) greatly benefited the community by releasing the model called Stable Diffusion, a latent diffusion model trained on the subset of the LAION-5B dataset (Schuhmann et al.). Since the shift from learning a task-specific model from scratch to leveraging pre-trained large-scale models is an inevitable trend in deep generative modeling, it is necessary to develop methods to better utilize these models.

To utilize the pre-trained models for generating the custom dataset, one can simply try to fine-tune the pre-trained generative model on a custom dataset, but this approach does not scale well with the model size. Another possible approach is to select the text prompt carefully by trial and error so that a text-to-image model generates the desired samples. This is possible but limited to datasets



(a) Prompt optimization

(b) Image generation

Figure 1: Overview of the proposed DYOD. (a) We first obtain the dataset representative text prompt that we will use to condition the reverse process by optimizing reconstruction loss. (b) Next, we guide the pre-trained text-to-image diffusion model by initializing its reverse process with either an image or the perturbed distribution of the custom generator. Image initialization can be used to transform a stroke, photo, etc. into the target domain, and distribution initialization is for unconditional synthesis.

that can be fully described with a single sentence. For example, with a text prompt *"A photo of a celebrity's face,"* a text-to-image model may reliably generate images that resemble the images of the CelebAHQ dataset (Karras et al., 2017). However, in most cases, it is difficult to model the dataset by solely adjusting a text prompt when the characteristics of the target dataset are not described in a single sentence. Moreover, text-to-image diffusion models such as Stable Diffusion tend to have difficulty in controlling the geometry of objects, such as locations, numbers, etc.

To address this, here we propose a method dubbed Diffusion model for Your Own Data (DYOD) that can effectively utilize a pre-trained text-to-image diffusion model to approximate the implicit distribution of a custom dataset. As shown in Fig. 1, DYOD is comprised of three steps. First, we obtain a dataset representative text prompt via optimizing the text embedding (Gal et al., 2022). As we directly explore the semantic latent space instead of the text space, the resulting text prompt can faithfully depict the characteristics of the dataset. Second, to generate the samples of the target distribution, we manipulate the implicit distribution of the diffusion generative prior by replacing the initial distribution of the reverse diffusion process with a learned one from user's data or an image initialization using user's stroke, photos, etc. Finally, by refining the initializations through the prompt conditioned pre-trained diffusion model, we can customize pre-trained diffusion generative models for our own data.

Our contributions are as follows:

- We present DYOD, a novel method of applying pre-trained text-to-image diffusion models for generative modeling to a custom dataset. This is enabled by guiding the reverse process of the diffusion model with the dataset representative text prompt and user-oriented initial distribution.
- In contrast to Gal et al. (2022) that mainly focuses on prompt engineering, we demonstrate the importance of initialization in customizing a pre-trained model to user-specific data set.
- Experimental results confirm that customized DYOD outperforms the Stable Diffusion baseline both qualitatively and quantitatively with accelerated sampling speed.

2 Background

Diffusion models (Ho et al., 2020) are hierarchical latent variable models defined as

$$p_\phi(\boldsymbol{x}_{0:T}) = p_\phi(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \tag{1}$$

where $\boldsymbol{x}_0 = \boldsymbol{x}$ is an observed variable we are interested in, $p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ is parameterized as a Gaussian distribution, and $p(\boldsymbol{x}_T)$ is set to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The joint distribution $p_\phi(\boldsymbol{x}_{0:T})$ is called the reverse process and can be conditioned on $\boldsymbol{c}$ for conditional generative modeling, i.e., $p_\phi(\boldsymbol{x}_{0:T}|\boldsymbol{c})$. Since a marginal density $p_\phi(\boldsymbol{x}) = \int p_\phi(\boldsymbol{x}_{0:T})d\boldsymbol{x}_{1:T}$ is intractable, we introduce a variational posterior, the forward process

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \tag{2}$$

where $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ is defined as

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}), \tag{3}$$

with a pre-defined noise schedule $\{\beta_t\}_{t=1}^N$. Notably, we can directly obtain $\boldsymbol{x}_t$ from $\boldsymbol{x}_0$ for an arbitrary t:

$$q_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{4}$$

with $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. For image synthesis, diffusion models are usually trained by minimizing the re-weighted variant of variational bound objective as follows:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}),\boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}),t\sim\mathcal{U}\{1,T\}}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t)||_2^2], \tag{5}$$

where $p(\boldsymbol{x})$ is a data distribution, $\boldsymbol{\epsilon}_\phi$ is a diffusion network, and $\boldsymbol{\epsilon}$ is a Gaussian noise added to the data to sample $\boldsymbol{x}_t$. As Eq. (4) allows us to efficiently sample $\boldsymbol{x}_t$ for any t, T can be an arbitrarily large number without sacrificing the tractability of training. In fact, diffusion models can even be extended into a continuous time in the limit of $T \to \infty$, where Song et al. (2020b) found an interesting connection between stochastic differential equations. After training, a reverse process of diffusion models are initialized into a standard Gaussian distribution, gradually removing the noise starting from $t = T$ to generate data:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t)\Big) + \sigma_t\bar{\boldsymbol{\epsilon}}, \tag{6}$$

where $\bar{\boldsymbol{\epsilon}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and σ_t is the standard deviation of the reverse process that can be either learned or fixed to constant.

3 Customizing Diffusion Models for Your Own Data

Since training a modern generative model from scratch is not easy with the limited computational resources and training data, our goal is to model the underlying distribution $\tilde{p}(\boldsymbol{x})$ of a custom dataset using a pre-trained text-to-image diffusion model $p_\phi(\boldsymbol{x}|\boldsymbol{c})$. In this section, we provide a detailed description of DYOD that is designed to address this issue. We first describe the procedure to obtain the dataset representative text prompt $\boldsymbol{c}^*$ via textual inversion (Gal et al., 2022) such that $p_\phi(\boldsymbol{x}|\boldsymbol{c}^*)$ is close to $\tilde{p}(\boldsymbol{x})$ (Sec 3.1). Next, we propose to further guide the diffusion model by adjusting the initial distribution of its reverse process (Sec 3.2). This procedure is summarized in Fig. 1.

3.1 Obtaining dataset representative text prompt

To apply $p_\phi(\boldsymbol{x}|\boldsymbol{c})$ to approximate $\tilde{p}(\boldsymbol{x})$, we first need to find $\boldsymbol{c}$ that brings $p_\phi(\boldsymbol{x}|\boldsymbol{c})$ as close as possible to $\tilde{p}(\boldsymbol{x})$. Although one can attempt to tune the text prompt manually, this is suboptimal as 1) searching is performed on the text space rather than the text embedding space, and 2) the quality of $\boldsymbol{c}$ is bounded by the user's proficiency with prompt engineering. Recently, Gal et al. (2022) proposed a novel textual inversion technique to find the word embedding v that represents the given 3-5 examples. We find that textual inversion can also be applied to a larger number of images, the entire dataset. To do so, we use the context texts to construct the text prompt $\boldsymbol{c}$ like "A photo of S_*," where S_* is a pseudo word that corresponds to v. Formally, we minimize the following objective to obtain the word embedding v^* that best represents the characteristics of $\tilde{p}(\boldsymbol{x})$:

$$v^* = \arg\min_v \mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}(\boldsymbol{x}),\epsilon\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}),t\sim\mathcal{U}\{1,T\}}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t, \boldsymbol{c}(v))||_2^2], \tag{7}$$

which is equivalent to find the word embedding v^* using reconstruction loss such that

$$v^* = \arg\min_v \mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}(\boldsymbol{x}),\epsilon\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}),t\sim\mathcal{U}\{1,T\}}\left\|\boldsymbol{x}_0 - \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t, \boldsymbol{c}(v))\right)\right\|_2^2, \tag{8}$$

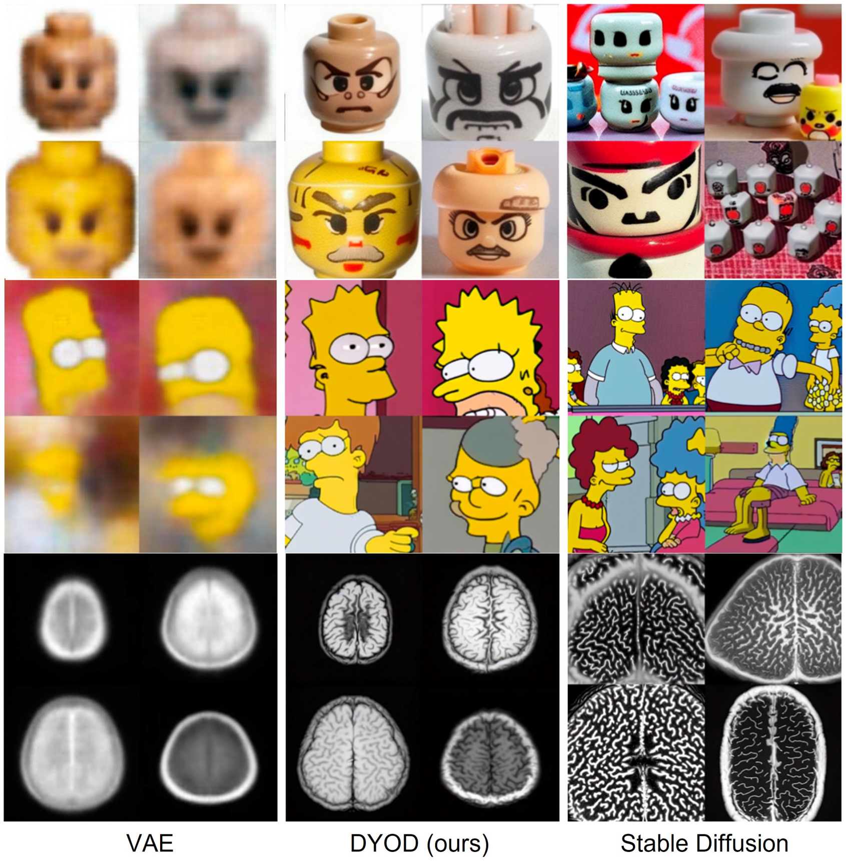


Figure 2: Generated samples by VAE, DYOD, and Stable Diffusion conditioned on $\boldsymbol{c}^*$. Samples from Stable Diffusion do not comply with the characteristics of the datasets, such as the number of objects, background, and object size (right).

where $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$. After finding v^*, we obtain the dataset representative text prompt $\boldsymbol{c}^*$ using v^* and the context texts. This procedure is illustrated in Fig. 1(a).

In addition to the fact that we provide a more systematic way rather than manually tuning the $\boldsymbol{c}$ by trial and error, exploring the semantic latent space allows us to find the $\boldsymbol{c}^*$ that is not obtainable by searching the text space.

3.2 Initializing the reverse process

Even though $\boldsymbol{c}^*$ represents the characteristics of the custom dataset, as shown in the rightmost panel of Fig. 2, Stable Diffusion with a text prompt alone cannot provide sufficient guidance to reliably model $\tilde{p}(\boldsymbol{x})$ in many cases. In particular, the number of the objects, locations, and other geometric configurations are difficult to control in Stable Diffusion. To settle the issue, we propose to guide the diffusion model by adjusting the initial distribution of its reverse process, as shown in Fig. 1(b), where two methods are considered: distribution initialization and image initialization.

Distribution initialization For the unconditional generation, we first train a small custom generative model on $\tilde{p}(\boldsymbol{x})$ and use its perturbed distribution as an initial distribution of the reverse process. Although our method is agnostic to the choice of the generative model, we train VAEs (Kingma & Welling, 2013) as they are easy to train and have good mode coverage. Furthermore, disentangling important geometric features are relatively easier in VAE and its variants.

Formally, we define our generative model as

$$p_{\theta,\phi,t_0|\boldsymbol{c}^*}(\boldsymbol{z},\boldsymbol{x}',\boldsymbol{x}_{0:t_0}) = p(\boldsymbol{z})p_\theta(\boldsymbol{x}'|\boldsymbol{z})q_{t_o|0}(\boldsymbol{x}_{t_0}|\boldsymbol{x}')p_\phi(\boldsymbol{x}_{0:t_0-1}|\boldsymbol{x}_{t_0},\boldsymbol{c}^*) \tag{9}$$

t_0	$0.3T$	$0.4T$	$0.5T$	$0.6T$	$0.7T$	$0.8T$	Stable Diffusion	VAE
Apple-orange	4.73	4.14	3.39	2.67	**2.40**	2.58	9.58	10.18
Lego-face	5.21	2.68	1.34	1.11	1.11	**1.03**	7.22	19.74
Simpsons-faces	29.52	26.02	18.88	7.90	1.94	**1.45**	3.65	37.48
Speed (T/t_0)	×3.3	×2.5	×2	×1.6	×1.4	×1.25	×1	-

Table 1: KID × 100 and sampling speed of DYOD. The results of Stable Diffusion and VAE are provided as a baseline.

with the prior distribution $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, VAE likelihood $p_\theta(\boldsymbol{x}'|\boldsymbol{z})$, noise distribution $q_{t_0|0}(\boldsymbol{x}_{t_0}|\boldsymbol{x}')$, and the reverse process $p_\phi(\boldsymbol{x}_{0:t_0-1}|\boldsymbol{x}_{t_0}, \boldsymbol{c}^*)$ starting from $t = t_0$. The marginal density of DYOD we are trying to integrate is as follows:

$$p_{\theta,\phi,t_0|\boldsymbol{c}^*}(\boldsymbol{x}) = \int p_{\theta,\phi,t_0|\boldsymbol{c}^*}(\boldsymbol{z}, \boldsymbol{x}', \boldsymbol{x}_{0:t_0}) d\boldsymbol{z} d\boldsymbol{x}' d\boldsymbol{x}_{1:t_0}. \tag{10}$$

It is noteworthy that after training a custom generative model, the sampling speed of DYOD is faster than standard diffusion models since we start the reverse process at $t_0 < T$ using the CCDF (come-closer-diffuse-faster) procedure in Chung et al. (2022).

Image initialization For the conditional synthesis, one can also initialize the reverse process with an initial image $\boldsymbol{y}$ that may come from stroke, photos, etc. Then, the learned $\boldsymbol{c}^*$ transforms an input image $\boldsymbol{y}$ into the sample from the target distribution $\tilde{p}(\boldsymbol{x})$. More specifically, for $t' \in (0, T]$, we perturb $\boldsymbol{y}$ by sampling $\boldsymbol{y}_{t'}$ from noise distribution $q_{t'|0}(\boldsymbol{y}_{t'}|\boldsymbol{y})$ and denoise it through the reverse process conditioned on $\boldsymbol{c}^*$. Formally, the transformed image $\boldsymbol{y}_0$ is sampled from the following distribution:

$$p_{\phi,t'|\boldsymbol{c}^*}(\boldsymbol{y}_0|\boldsymbol{y}) = \int q_{t'|0}(\boldsymbol{y}_{t'}|\boldsymbol{y}) p_\phi(\boldsymbol{y}_{0:t'-1}|\boldsymbol{y}_{t'}, \boldsymbol{c}^*) d\boldsymbol{y}_{1:t'}, \tag{11}$$

where $p_\phi(\boldsymbol{y}_{0:t'-1}|\boldsymbol{y}_{t'}, \boldsymbol{c}^*)$ denotes a reverse process starting from $t = t'$.

4 Experiments

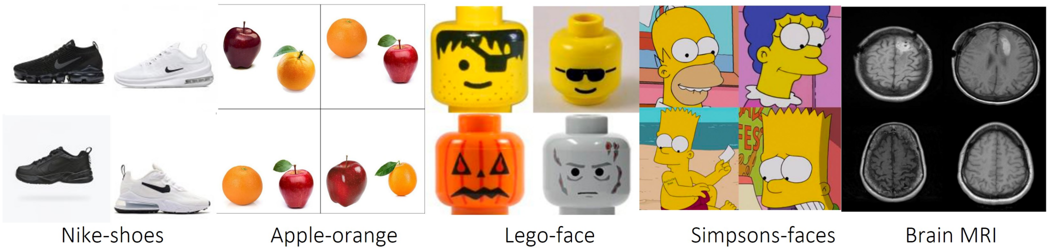


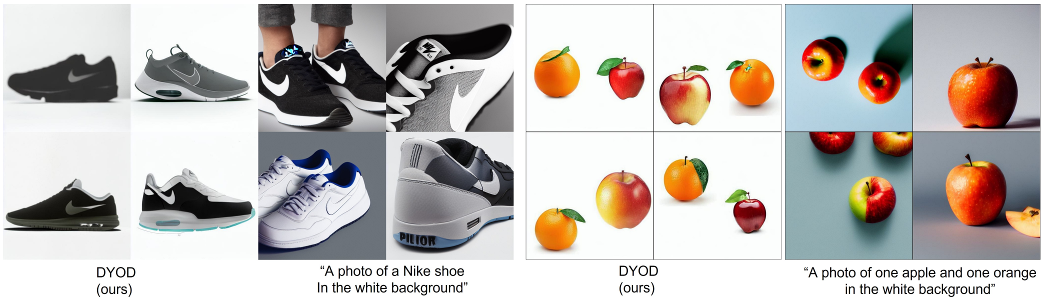
Figure 3: Examples of the datasets we use in our experiments.

4.1 Experimental Setup

We evaluate our method on five datasets. **Nike-shoes** dataset contains approximately 200 128 × 128 Nike shoe images from the web. **Apple-orange** dataset consists of 30,000 512 × 512 images synthesized by randomly placing an apple and an orange image on a white background. **Lego-face**[1] dataset contains approximately 2,600 128 × 128 images of the faces of the lego figures. **Simpsons-faces**[2] dataset contains approximately 10,000 200 × 200 images of the faces of The Simpsons (Wikipedia, 2022) characters that are extracted from the video files. **Brain MRI** is a subset of fastMRI (Zbontar

[1] https://github.com/iechevarria/lego-face-VAE

[2] https://www.kaggle.com/datasets/kostastokis/simpsons-faces



DYOD (ours)

"A photo of a Nike shoe In the white background"

DYOD (ours)

"A photo of one apple and one orange in the white background"

Figure 4: Comparison with Stable Diffusion on Nike-shoes and Apple-orange datasets.

et al., 2018) dataset consisting of 30,000 images that are resized to 256 × 256 resolution. See Fig. 3 for the examples of each dataset.

We note that our aim is not to pursue state-of-the-art performance on the standard benchmark datasets but to enable users to build the generative model on their custom datasets that are difficult to generate by simply tuning the text prompt. So, throughout our experiments, we limit our computational resources to a single Geforce 1080 Ti. For the text-to-image diffusion model, we use the publicly available Stable Diffusion model. We use DDIM (Song et al., 2020a) sampler with $T = 50$. We employ the vanilla VAE implementation by Subramanian (2020) as it can be trained efficiently. Training VAE takes 3 minutes, 3 minutes, 30 minutes, 30 minutes, and an hour on the Nike-Shoes, Lego-face, Apple-orange, Simpsons-faces, and Brain MRI datasets, respectively. It is obvious that using the modern hierarchical VAEs (Child, 2020; Vahdat & Kautz, 2020) or other types of generative models will improve performance, but we leave it for future work. Since Stable Diffusion does not support resolutions other than 512×512, we upsample the input images smaller than 512×512 before feeding into the Stable Diffusion and then downsample into the original resolution. We use Lanczos interpolation, and we find that using the upsampling neural network (Wang et al., 2021) does not yield better results.

To measure the perceptual similarity between two distributions, we use Kernel Inception Distance (KID) (Bińkowski et al., 2018), where the squared Maximum Mean Discrepancy is measured between two sets of the feature vectors.

4.2 Experimental results on initialization

As shown in Fig. 4, Stable Diffusion fails to synthesize the images with the correct object size and viewpoints on Nike-shoes and Apple-orange datasets. Moreover, the generated images by Stable Diffusion do not follow the text prompt, resulting in the incorrect number of objects or object type. The results demonstrate that a text prompt does not provide sufficient guidance to reliably model the target distribution. By guiding the reverse process with the custom generator that learned the statistics of the target distribution, our DYOD effectively applies the knowledge obtained from the custom dataset to diffusion prior.

Fig. 2 demonstrates that DYOD successfully generates the samples of the target datasets. Contrarily, the results of Stable Diffusion appear to contain some concepts from the datasets, but they do not comply with important characteristics such as the number of objects, background, and object size, despite using the optimized prompt. Therefore, we can conclude that it is vital to guide the diffusion model not only by using a proper text prompt but by adjusting the initial distribution of the reverse process. Table 1 demonstrates that our method significantly outperforms the Stable Diffusion in KID (Bińkowski et al., 2018) for both datasets. Compared to Stable Diffusion baseline, which corresponds to our method with $t_0 = T$, DYOD can generate more high-quality samples with accelerated speed. Notably, DYOD achieves lower KID than Stable Diffusion while generating samples up to 3.3 times faster.

Image initialization As shown in Fig. 5, the dataset representative text prompt allows DYOD to effectively transform the input images such as stroke or photos into the images of the Lego-face and

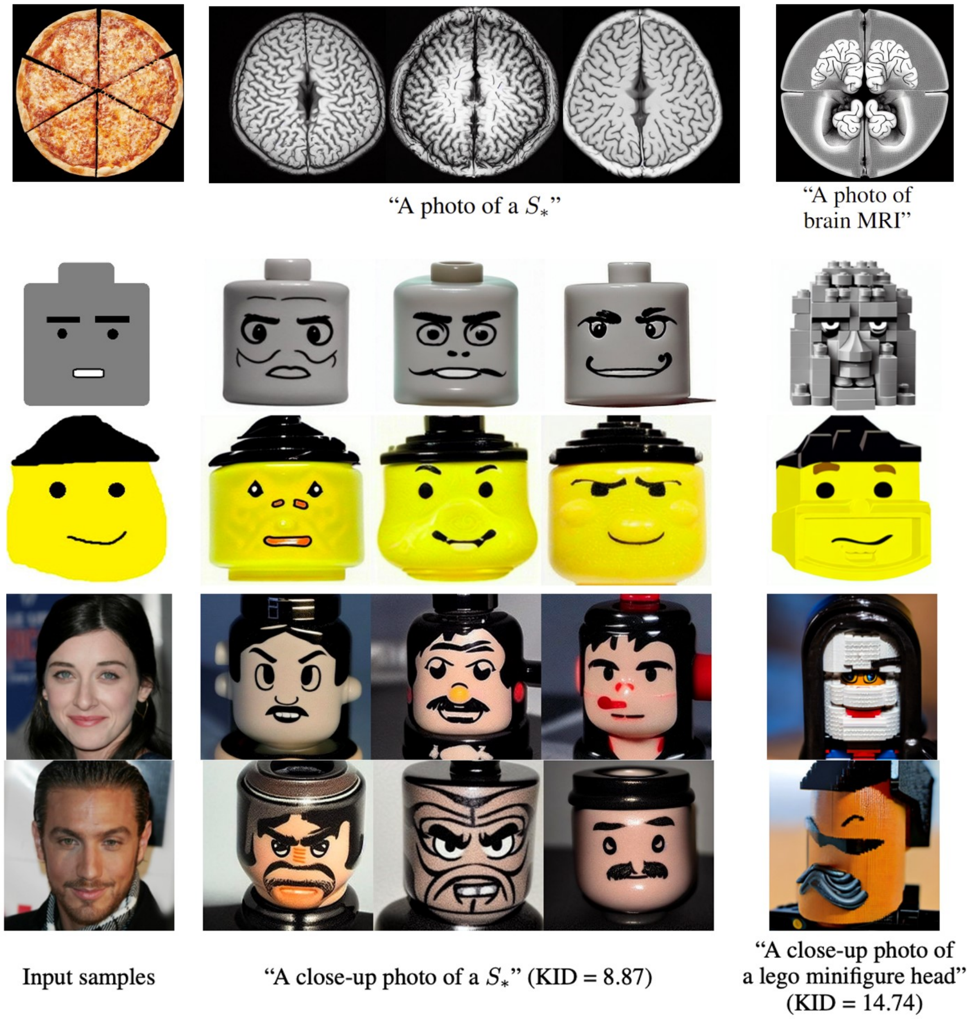

"A photo of a S_*"
"A photo of brain MRI"
Input samples
"A close-up photo of a S_*" (KID = 8.87)
"A close-up photo of a lego minifigure head" (KID = 14.74)

Figure 5: Image initialization results. The results of Pizza-to-BrainMRI are on the top, Stroke-to-Lego on the middle, and CelebAHQ-to-Lego on the bottom. For Pizza-to-BrainMRI, the input sample is converted to grayscale before being fed to the model. Using the dataset representative text prompt, our DYOD successfully transforms the input image into the samples of the target dataset. Without using the dataset representative text prompt, a text prompt is not sufficient for Stable Diffusion to depict the characteristics of the dataset, resulting in a different appearance compared to the custom datasets (see Fig. 3). We also report KID × 100 for CelebAHQ-to-Lego results.

brain MRI data set, respectively. Without using the dataset representative text prompt, the generated samples from Stable Diffusion do not resemble the Lego-face and brain MRI data, resulting in higher KID (see the rightmost panel of Fig. 5).

4.3 Experimental results on flexible latent manipulation

Latent space interpolation To investigate the learned latent space of DYOD, we linearly interpolate between two VAE latent variables, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, with the interpolation parameter $\lambda \in [0, 1]$, i.e., $\boldsymbol{z} = \boldsymbol{z}_1 + \lambda(\boldsymbol{z}_2 - \boldsymbol{z}_1)$. As we employ the deterministic reverse process, the remaining sources of randomness are the VAE latent variable $\boldsymbol{z}$ and additive Gaussian noise $\boldsymbol{\epsilon}$. Fig. 6 shows that without fixing $\boldsymbol{\epsilon}$, the properties such as facial expressions and beard and eyebrow shapes vary among interpolated images. Fixing $\boldsymbol{\epsilon}$ results in a smooth interpolation, demonstrating that DYOD learns meaningful latent space.

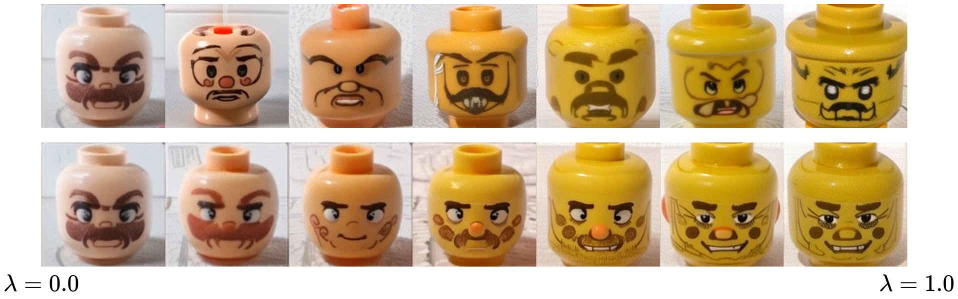


Figure 6: Linear interpolation in the latent space of VAE. λ denotes the interpolation factor. *Top row*: DYOD generated samples without fixing ϵ. *Bottom row*: DYOD generated samples with fixed ϵ, resulting in smooth interpolation.

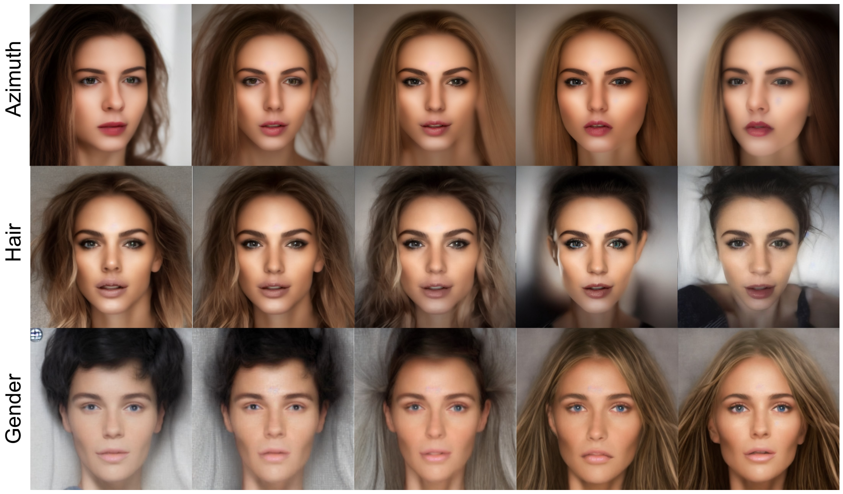


Figure 7: Manipulating the VAE latent variable of DYOD while other stochasticities are controlled. We traverse a single dimension of latent vectors while keeping others fixed.

Disentangled latent manipulation One drawback of diffusion models is that they do not provide the compact latent representation of data. Contrarily, thanks to the disentangled latent space of VAE, we can manipulate the high-level attributes of images by controlling the latent vector of DYOD. As shown in Fig. 7, DYOD successfully controls the azimuth, hair length, and gender of the generated images.

5 Related Works

Prior to our work, several studies utilized diffusion prior to refine the images. Meng et al. (2021) first presented an image editing method based on diffusion models where they perturb an input stroke and transform it into a realistic image. In Chung et al. (2022), the authors refine the neural network prediction using the pre-trained diffusion model, but in a different context of conditional synthesis. Although these two studies employed the pre-trained diffusion model, they assumed that the unconditional diffusion model is trained on a class-specific dataset (e.g., LSUN, FFHQ, AFHQ, etc) and did not consider the method for guiding the text-conditional diffusion prior trained on the more diverse dataset. Ryu & Ye (2022) recently apply diffusion prior to increase the resolution of previous samples in their coarse-to-fine generation scheme. This particular work also can be seen as learning the initial distribution of the reverse process, but their method requires training a diffusion model from scratch. In Pandey et al. (2022), the authors discuss the symbiosis of VAEs and diffusion models. They first train a VAE and subsequently a diffusion model conditioned on the VAE samples, which can be viewed as a generator-refiner framework. We note that our work is in a distinct direction with a different goal, as our aim is to utilize the *pre-trained* diffusion model by guiding its reverse process.

6 Limitations and Conclusion

In this paper, we introduced DYOD, a method to borrow the power of a pre-trained text-to-image diffusion model for generative modeling on a custom dataset. We proposed to guide the diffusion prior with the dataset representative text prompt and a better initial distribution of the reverse process. We showed that DYOD outperforms Stable Diffusion baseline by a large margin and has a faster sampling speed while being trained within a few hours with a single GPU.

A limitation of DYOD is that its performance is bounded by textual inversion. It is hard to find the dataset representative text prompt if the dataset is not in the support of the diffusion prior $p_\phi(\boldsymbol{x}|\boldsymbol{c})$, although we believe that this problem may be reduced as the scale of pre-trained models increases. Fine-tuning the ϕ as Ruiz et al. (2022) might be able to settle the issue at the expense of increased computational costs, which we leave for future work.

References

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12413–12422, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

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.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

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

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

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*, 2022.

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

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*, pp. 10684–10695, 2022.

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. *arXiv preprint arXiv:2208.12242*, 2022.

Dohoon Ryu and Jong Chul Ye. Pyramidal denoising diffusion probabilistic models. *arXiv preprint arXiv:2208.01864*, 2022.

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

Christoph Schuhmann, Romain Beaumont, Cade W Gordon, Ross Wightman, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, et al. Laion-5b: An open large-scale dataset for training next generation image-text models.

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

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

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

A.K Subramanian. Pytorch-vae. `https://github.com/AntixK/PyTorch-VAE`, 2020.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1905–1914, 2021.

Wikipedia. The Simpsons — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=The%20Simpsons&oldid=1109791430`, 2022. [Online; accessed 21-September-2022].

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.