ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10445885

PROGRESSIVE IMAGE SYNTHESIS FROM SEMANTICS TO DETAILS WITH DENOISING DIFFUSION GAN

Guoxing Yang[†], Haoyu Lu[†], Chongxuan Li[†], Guang Zhou[⋆], Haoran Wu[⋆], Zhiwu Lu[†]*

[†] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[⋆] China Unicom Research Institute, Beijing, China

ABSTRACT

Although denoising diffusion probabilistic models (DDPMs) have shown remarkable progress in image generation, they typically face two main challenges: the time-expensive sampling process and the semantically meaningless latent space, which are often addressed separately in previous works. In particular, the latest representative work Denoising Diffusion GAN reduces the sampling steps to as few as two but ignores the semantics of the latent space. To address the two challenges simultaneously, we propose a two-stage framework to make the latent space of Denoising Diffusion GAN more semantically meaningful while enjoying its efficiency. Extensive results on three benchmark datasets demonstrate that our proposed diffusion model achieves competitive results with only two sampling steps in unconditional image generation. More importantly, the latent space of our diffusion model trained for unconditional image generation is shown to be semantically meaningful, which can be exploited on various downstream tasks (e.g., attribute editing) without further training.

Index Terms— Diffusion model, GAN, Latent space, Semantics

1. INTRODUCTION

Image generation falls in the most popular topics in computer vision, which has been dominated by Generative Adversarial Networks (GANs) [1] in the past few years [2, 3, 4] due to their superior ability to synthesize photo-realistic images. Recently, Denoising Diffusion Probabilistic Models (DDPMs) [5, 6, 7, 8] have achieved impressive results in image generation [9, 10, 11], which are shown to outperform GANs in terms of sample quality, diversity, and training stability.

Although DDPMs show superior ability in generation tasks, they typically face two main drawbacks (but GANs do not): the time-expensive sampling process and the semantically meaningless latent space, which significantly limit their applications in practice. Existing works [12, 13, 14] have started to address these two challenges independently, but it is still a dilemma in the literature. In particular, Denoising Diffusion GAN [15] (DDGAN for short) takes as few as two sampling steps but achieves competitive sample quality and diversity w.r.t. the traditional DDPMs. However, it ignores the semantics of the latent space.

In this paper, we propose to decompose the sampling process of DDPMs into two stages: a semantics generation stage and a detail refinement stage, and devise a two-stage framework to enhance the semantics of the latent space in DDGAN while enjoying its efficiency simultaneously. Specifically, in the semantics generation stage, we introduce a semantic encoder to encode the input image into a latent vector and enforce the generator to recover the corresponding semantics of the input image from the pure Gaussian noise conditioned on the latent vector. In the detail refinement stage, we encourage the generator to refine the details while preserving the main semantic information of the output of the semantics generation stage. With such a two-stage framework, the semantics of generated images is mainly controlled by the latent vectors derived from the semantic encoder, and the corresponding latent space becomes semantically meaningful. Furthermore, instead of sampling the time step uniformly during training, we start by sampling the semantics generation stage more frequently and end by sampling the detail refinement stage more frequently with a linear scheduler. This adaptive strategy further improves the sample quality and enhances the semantics of the latent space of our diffusion model.

Our main contributions are three-fold: **(1)** To the best of our knowledge, this is the first work to address the two main challenges of DDPMs simultaneously, which allows that a unconditional DDPM can be directly deployed for various downstream tasks with efficient sampling process. **(2)** Based on DDGAN, we provide a simple yet effective two-stage framework to enhance the semantic of the latent space of this model while enjoying its sampling efficiency. Our newly-devised progressive training pipeline further improves the sample quality as well as enhance the semantics of the latent space. **(3)** Extensive results on three benchmark datasets show that our diffusion model achieves competitive performance but with only two sampling steps in unconditional image generation. Importantly, the latent space of our model trained for unconditional generation is shown to be semantically meaningful, which can be exploited on various downstream tasks (e.g., attribute editing) without further training.

2. PROGRESSIVE DENOISING DIFFUSION GAN

2.1. Proposed Method

With careful exploration of the two-step DDGAN, we empirically find that the semantics of its generated images mainly depends on the latent variables inputted in the first sampling step. As a result, the sampling process of DDGAN [15] can be decomposed into two stages: the semantics generation stage and the detail refinement stage. Based on this two-stage sampling process, our goal is to make the latent spaces in the semantics generation stage more semantically meaningful, so that we can address the two main drawbacks of DDPMs (i.e., the time-expensive sampling process and the semantically meaningless latent space) simultaneously. To achieve this, we present a two-stage framework in the following (see Figure 1). For easier understanding, we denote the image sampled from $q(x_{1:t}|x_0)$ and $p_\theta(x_{t:T})$ as $\bar{x}_t$ and x_t, respectively.

Semantics Generation Stage. As mentioned above, we should make the latent variables inputted in this stage more semantically meaningful. To this end, we design an auxiliary encoder that learns to encode the input image $\bar{x}_0$ into a latent vector z_2, and adopt the generator to synthesize images conditioned on z_2.

[*] Corresponding author

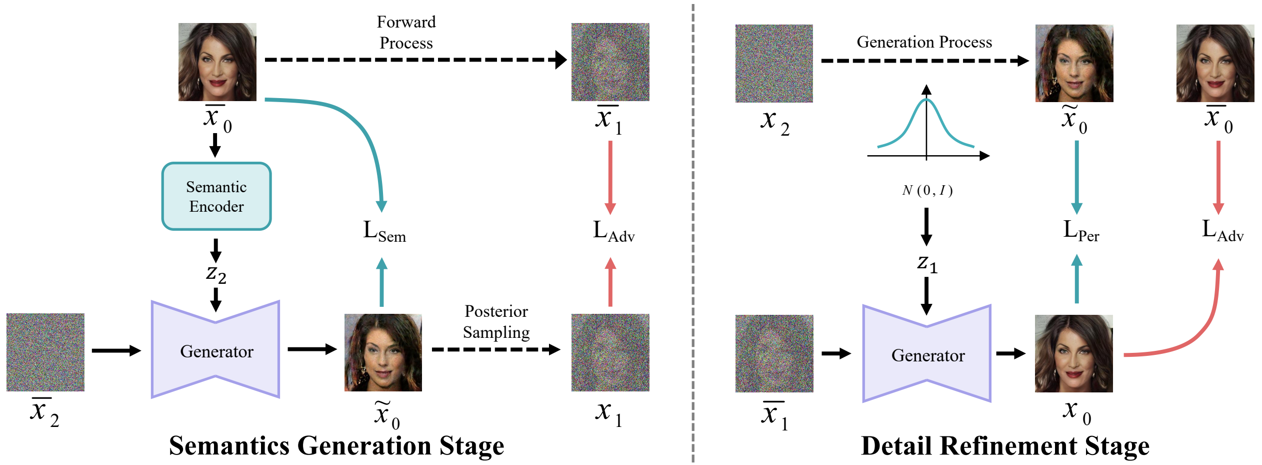

Fig. 1. Illustration of the training process of our proposed model. $\bar{x}_t$ and x_t denote images sampled frome $q(x_{1:t}|x_0)$ and $p_\theta(x_{t:T})$, respectively. Note that the two novel losses L_{Sem} and L_{Per} are essential to learn semantically meaningful latent space.

The similar idea of introducing an auxiliary encoder to learn a semantic latent space of DDPMs is also explored in Diffusion Autoencoders [14] (DiffAE for short), which is trained by minimizing the loss function $\|\epsilon_\theta(x_t, t, z_{sem}) - \epsilon\|_2^2$ based on the Gaussian assumption. However, when the Gaussian assumption is removed for fast sampling in DDGAN, directly applying DiffAE to DDGAN *is found to be invalid* (see Sec. 3.2). To address this challenge, we thus carefully explore which component indeed makes DiffAE valid, and find that the success of DiffAE fundamentally attributes to its special form of loss function (i.e., L2 loss), which implicitly enforces x_t to be close to x_0 in semantics. However, when the Gaussian assumption is removed, the adversarial loss is adopted for training instead of the L2 loss. Further, the adversarial loss can only guarantee the two distribution $q(x_{t-1}|x_t)$ and $p_\theta(x_{t-1}|x_t)$ to be close, but can not guarantee x_t and x_0 to be close in semantics. Overall, the use of adversarial loss leads to the failure of simply applying DiffAE to DDGAN.

In this work, we thus propose a novel perceptual loss L_{Per} to address the challenge above. Concretely, we apply the L1 constraint to the input image $\bar{x}_0$ and the output of the generator $\tilde{x}_0$ at both pixel level and feature level to guarantee that they are close in semantics:

$$\begin{aligned}\mathcal{L}_{Sem} = \mathbb{E}_{q(x_0)q(x_2|x_0)}\big[&\|G(\bar{x}_2, SE(\bar{x}_0), t=2) - \bar{x}_0\|_1 \\ &+ \|V(G(\bar{x}_2, SE(\bar{x}_0), t=2)) - V(\bar{x}_0)\|_1\big],\end{aligned} \tag{1}$$

where $V(\cdot)$ is to extract feature with the pre-trained VGG [16].

With such framework and loss, z_2 is encouraged to represent high-level semantic information, while x_2 is encouraged to represent low-level detail information. The training process of the semantics generation stage is shown in Figure 1 (left).

Detail Refinement Stage. With the output of the semantics generation stage, the goal of detail refinement stage is to refine its details while preserving the semantic information of it. Although the latent variables inputted in this stage (i.e., z_1) affects only imperceptible details in DDGAN, we find that it leads to much semantics variation in our framework. Therefore, to further guarantee z_1 has no influence on semantics of generated images, we additionally apply the perceptual constraint L_{Per} to the output of generator and $\tilde{x}_0$ in this stage. Formally, the perceptual loss L_{Per} is defined as follows:

$$\mathcal{L}_{Per} = \mathbb{E}_{q(x_0)q(x_t|x_0)p_\theta(x_{t+1:T})}\big[\|G(\bar{x}_1, z_1, t=1) - \tilde{x}_0\|_1\big]. \tag{2}$$

With this constraint, the semantics of generated images are controlled by z_2, and z_1 only affects the imperceptible details, which is important for attribute manipulation and controllable generation (see Figure 2). The training process of this stage is shown in Figure 1 (right).

Progressive Training. Although the sampling process of DDGAN has been decomposed into two stages in the above formulations, we still train our model end-to-end like traditional DDPMs, instead of separating the training into two distinct phases. For traditional DDPMs, each denoising step is independent to the other steps during training, and they thus randomly sample t from the uniform distribution in each training step. However, in the detail refinement stage, the training of generator G depends on the output of the semantics generation stage $\tilde{x}_0$, which can not provide precise semantic information at the beginning of training due to the inaccurate sampling process $p_\theta(x_{1:2})$. To alleviate this issue, we present a novel progressive training pipeline, which progressively transforms the attention from semantics to sample quality. Concretely, in each training step, we pick $t = 2$ (i.e., the semantics generation stage) with the probability P and $t = 1$ (i.e., the detail refinement stage) with the probability $1 - P$. P linearly decreases as the number of training epochs increases:

$$P = P_{max} - (P_{max} - P_{min})\frac{n_e}{N_e}, \tag{3}$$

where n_e and N_e denote the current and total number of training epochs, respectively. P_{max} and P_{min} are two hyperparameters that control the trade-off between the semantics and sample quality. With such dynamic sampling strategy, the model pays more attention to the semantics generation at the beginning of training. As the training process goes on, the model gradually diverts attention to the detail refinement. When this happens, the model has learned to generate semantically meaningful image $\tilde{x}_0$, and thus the refinement steps can be trained efficiently based on $\tilde{x}_0$.

3. EXPERIMENTS

3.1. Datasets and Settings

Datasets. To evaluate the effectiveness of our proposed diffusion model, we mainly conduct experiments on CelebA-HQ [2]. Moreover, we also conduct experiments on FFHQ [3] and LSUN-Churches [17], which are widely used to evaluate the sample quality of generative models. Images are resized to 256 × 256 in all of our experiments.

Evaluation Metrics. We adopt the Frechét Inception Distance (FID) [18] to evaluate the quality of generated images, which is commonly used in previous works. Following previous works [3, 14], we adopt the Perceptual Path Length (PPL) [3] to evaluate how the latent space of a generative model is semantically meaningful. Note that we follow them to set the ε to 1e-4 and divide the resultant PPL by ε^2. We generate 10,000 samples to evaluate PPL on all datasets.

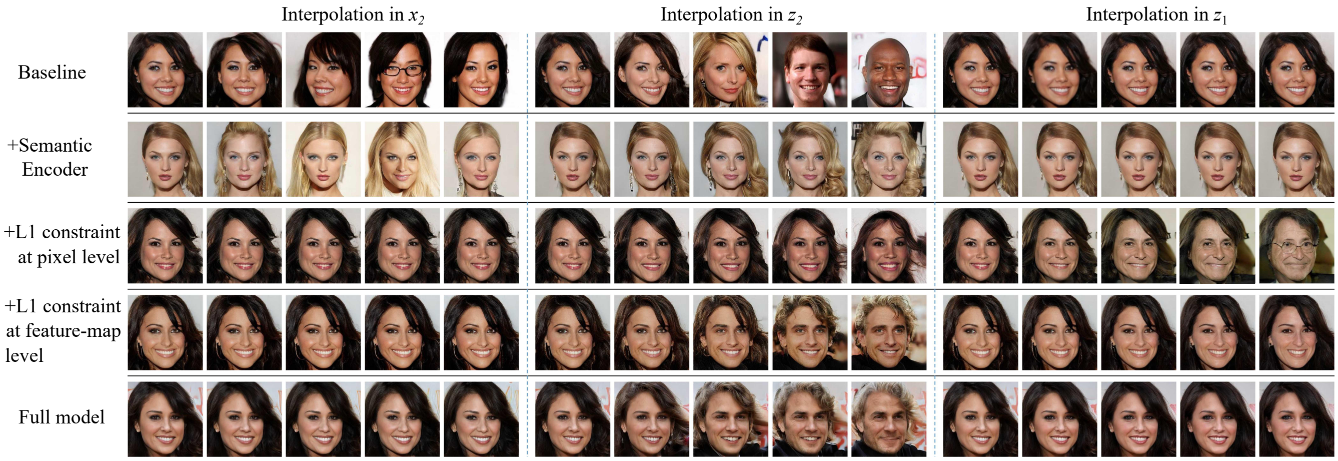

Fig. 2. Interpolation results on CelebA-HQ in different latent spaces of our ablated models. The semantics of images generated by our full model is mainly controlled by z_2, and its interpolation results are smooth, indicating that the latent space is **semantically meaningful**.

Table 1. The quantitative results of the ablation study on CelebA-HQ. The result in each row is obtained by adding the corresponding component to the model in the last row (except the first row). The second-best result is marked by underline. We empirically find that a variable leads to little semantics variation when the corresponding PPL< 50 (marked in gray).

Method	FID ↓	PPL ↓		
		x_2	z_2	z_1
Baseline	7.64	2135.71	8312.27	6.84
+ Semantic encoder	**5.88**	11649.05	7195.09	4.70
+ L1 constraint at pixel level	6.77	26.96	1576.84	185.31
+ L1 constraint at feature-map level	6.66	39.21	723.97	40.85
+ Progressive training pipeline (ours)	6.47	36.51	**640.81**	39.74

When comparing with the state-of-the-art methods, we only consider the PPL of z_2 for our model, which controls the main semantics of generated images. To evaluate the efficiency of sampling process, we also report the clock time of generating a batch of 100 images on a A100 GPU and the number of function evaluations (NFE) as metrics.

Table 2. Quantitative results on the CelebA-HQ dataset. The second-best result is marked by underline. The VAE and GAN based methods are marked in gray.

Method	FID ↓	PPL ↓	NFE ↓	Time (s) ↓
VQ-GAN [19]	10.20	-	257	12.96
VAEBM [20]	20.40	-	24	36.64
NVAE [21]	29.70	449.2	1	1.93
NCP-VAE [22]	24.80	-	1	-
DC-AE [23]	15.80	-	1	0.23
ADA [24]	15.21	178.9	1	0.77
Score SDE [13]	7.23	-	2000	8875.00
UDM [25]	7.16	-	2000	-
P2 [26]	6.91	-	500	-
LDM [27]	**5.11**	9477.9	500	273.00
LSGM [28]	7.22	-	147	28.37
LDM [27]	20.58	9638.4	50	25.00
DiffAE [14]	15.76	845.6	50	55.70
DDGAN [15]	7.64	8312.3	**2**	**1.02**
Ours	6.47	**640.8**	**2**	**1.02**

3.2. Ablation Study

To demonstrate the contribution of each component of our proposed model, we conduct ablation study on the CelebA-HQ dataset. Concretely, we consider DDGAN [15] as the baseline, and add various components on the top of it gradually. We first add the semantic encoder (denoted as '+ Semantic encoder') on the top of the baseline, which injects the latent vector extracted from semantic encoder to the generator at each step instead of the random noises. Note that the resultant model can be considered as the simple combination of DDGAN [15] and DiffAE [14]. Further, we add the L1 constraint at pixel level to the model (denoted as '+ L1 constraint at pixel level'). Subsequently, we add the L1 constraint to the feature maps extracted by the VGG network when applying the two L1 constraints to the semantics generation stage only (denoted as '+ L1 constraint at feature-map level'). Finally, we add the constraint in Eq. (2) to the model and train it with the proposed progressive training pipeline (denoted as '+ progressive training pipeline').

Quantitative Results. The quantitative results of ablation study are shown in Table 1. It can be observed that: **(1)** The simple combination of DDGAN and DiffAE can bring a boost in FID. However, the PPLs of x_2 and z_2 are still very large, indicating that such simple combination fails to enhance the semantics of the latent space of the resultant model due to the gap between Gaussian assumption and non-Gaussian assumption. **(2)** Adding the L1 constraint at pixel level makes the semantics be mainly affected by z_2. Further, the space of z_2 is indeed semantically meaningful (see the PPL of z_2), indicating that this loss is essential to learn the semantics of the latent space. **(3)** Based on the findings mentioned in Sec. 2.1, we apply the above constraint to the semantics generation stage only, and additionally apply the L1 constraint at feature-map level, which brings a boost in both sample quality and semantics of latent space (see the FID and PPL of z_2). **(4)** Our progressive training pipeline further improves the FID and PPL of z_2, indicating that it can improve the sample quality as well as enhance the semantics of the latent space.

Qualitative Results. We further give qualitative analysis to investigate the effectiveness of our proposed components under human perception. Concretely, we explore the interpolation results in different spaces of our ablated models. As stated by previous works [29, 30], the interpolation results are more smooth, the latent space is more semantically meaningful. Firstly, we use the baseline model to randomly generate an image denoted as (x_2, z_2, z_1). We then separately move each variable along a random direction to generate a sequence of images. For the other ablated models, we adopt the semantic encoder of them to encoder (x_2, z_2, z_1) into the latent vector $\tilde{z}_2$, and generate an initial image denoted as $(\tilde{x}_2, \tilde{z}_2, \tilde{z}_1)$. We then generate in-

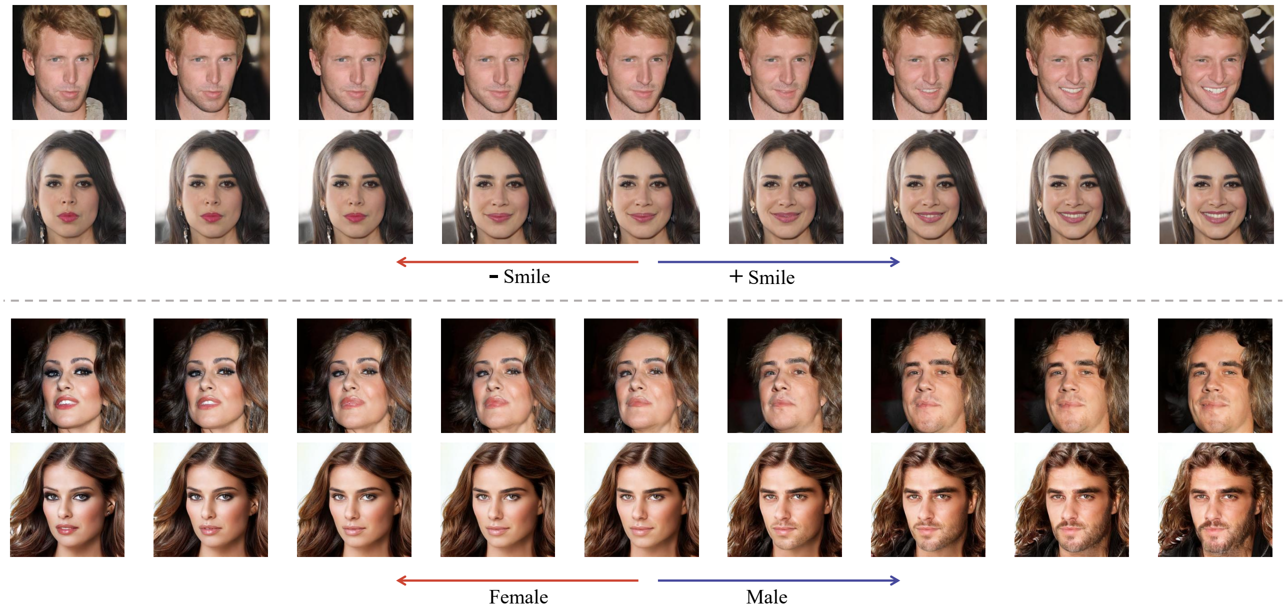

Fig. 3. Attribute manipulation results of our model on CelebA-HQ. We manipulate the local attribute **Smile** (the top two rows) and the global attribute **Gender** (the bottom two rows) by moving z_2 along the direction found by trained SVMs. Note that our model is trained for unconditional generation only, and is deployed for attribute manipulation without further training.

terpolation results in different latent spaces of each ablated model like the baseline model. The interpolation results are shown in Figure 2. We can observe that: **(1)** The interpolation results of the baseline are not smooth in the latent space of x_2, z_2, indicating that these two spaces of the baseline are semantically meaningless. In addition, z_1 only affects the imperceptible details of the generated images. **(2)** Adding the semantic encoder on the top of the baseline fails to make the latent space of x_2, z_2 more semantically meaningful, indicating that directly applying DiffAE to the non-Gaussian assumption is invalid due to the gap between non-Gaussian assumption and Gaussian assumption. **(3)** Adding the L1 constraint at pixel level makes the model can reconstruct the input images. In addition, the interpolation results in the space of z_2 are smooth, showing that this space is semantically meaningful. However, z_1 also affects some semantics of generated images (e.g., age and eye glasses). **(4)** Adding the L1 constraint at feature-map level (only applied to step T) further improves the smoothness of the latent space of z_2. Note that the semantics of generated images is mainly affected by z_2, but x_2 and z_1 still lead to little semantic variation (e.g., hair and identity). **(5)** Our proposed progressive training pipeline further guarantee that x_2 and z_1 affect only the imperceptible details of the generated images. Furthermore, both the sample quality and the smoothness of the latent space of z_2 are improved.

3.3. Comparison to the State-of-the-Arts

In this section, we compare our diffusion model with the state-of-the-arts in the unconditional generation tasks. Note that we only consider the PPL of z_2 for our model, since it controls the main semantics of generated images (see Sec. 3.2). The quantitative results on CelebA-HQ are shown in Table 2. We can observe that: **(1)** Our diffusion model outperforms all the competitors except LDM [27] in sample quality (i.e., FID). However, LDM requires 500 steps in sampling process, which takes **273** seconds to generate a batch of 100 images on a A100 GPU (but our model takes only **1.02** seconds). **(2)** Our model outperforms all DDPM based competitors on PPL, which indicates that the latent space of our model is more semantically meaningful. Note that the PPL of our model is even comparable to those of GANs and VAEs. **(3)** Our model also has a faster sampling process than the other DDPM based competitors (see NFE and Time in Table 2). **(4)** Particularly, our model beats DiffAE on all metrics, indicating that our model can generate more realistic images with much fewer sampling steps (i.e., much faster sampling process). Meanwhile, the latent space of our model is more semantically meaningful.

3.4. Attribute Manipulation

To further demonstrate that the latent space of our model is semantically meaningful, we apply our model trained for unconditional generation to attribute manipulation without further training. The details of conducting such attribute manipulation are given in Appendix. The continuous manipulation results by our model are shown in Figure 3. We find that our model can smoothly change both local and global attributes, while preserving the other information of original images, which indicates that the latent space of our model is semantically meaningful and disentangled.

4. CONCLUSION

In this work, we explore the latent space of DDGAN, and propose to decompose the sampling process into two stages motivated by the empirical findings from our exploration. With this two-stage sampling process, we propose a novel progressive training pipeline to address the two main challenges of DDPMs simultaneously, which are only separately explored in previous works. Extensive experiments show that our proposed model can achieve competitive results with only two sampling steps on unconditional images generation. Importantly, our diffusion model can generate smooth interpolation results and can be adopted in attribute manipulation without further training, indicating that the latent space of our model is semantically meaningful.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (62376274), and China Unicom Innovation Ecological Cooperation Plan.

5. REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NeurIPS 2014*, 2014, pp. 2672–2680.

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *ICLR*, 2018.

[3] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410.

[4] Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov, "Image generators with conditionally-independent pixel synthesis," in *CVPR*, 2021, pp. 14278–14287.

[5] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015, vol. 37, pp. 2256–2265.

[6] Yang Song and Stefano Ermon, "Generative modeling by estimating gradients of the data distribution," in *NeurIPS*, 2019, pp. 11895–11907.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020, pp. 6840–6851.

[8] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon, "ILVR: Conditioning method for denoising diffusion probabilistic models," in *ICCV*, 2021, pp. 14347–14356.

[9] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, pp. 47:1–47:33, 2022.

[10] Alexander Quinn Nichol and Prafulla Dhariwal, "Improved denoising diffusion probabilistic models," in *ICML*, 2021, vol. 139, pp. 8162–8171.

[11] Prafulla Dhariwal and Alexander Quinn Nichol, "Diffusion models beat GANs on image synthesis," in *NeurIPS*, 2021, pp. 8780–8794.

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021.

[13] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, "Score-based generative modeling through stochastic differential equations," in *ICLR*, 2021.

[14] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," *arXiv preprint arXiv:2111.15640*, 2021.

[15] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," in *ICLR*, 2022.

[16] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao, "LSUN: construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *NeurIPS*, 2017, pp. 6626–6637.

[19] Patrick Esser, Robin Rombach, and Björn Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021, pp. 12873–12883.

[20] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat, "VAEBM: A symbiosis between variational autoencoders and energy-based models," in *ICLR*, 2021.

[21] Arash Vahdat and Jan Kautz, "NVAE: A deep hierarchical variational autoencoder," in *NeurIPS*, 2020.

[22] Jyoti Aneja, Alexander G. Schwing, Jan Kautz, and Arash Vahdat, "A contrastive learning approach for training variational autoencoder priors," in *NeurIPS*, 2021, pp. 480–493.

[23] Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu, "Dual contradistinctive generative autoencoder," in *CVPR*, 2021, pp. 823–832.

[24] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila, "Training generative adversarial networks with limited data," in *NeurIPS*, 2020.

[25] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon, "Score matching model for unbounded data score," *arXiv preprint arXiv:2106.05527*, 2021.

[26] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon, "Perception prioritized training of diffusion models," *arXiv preprint arXiv:2204.00227*, 2022.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," *arXiv preprint arXiv:2112.10752*, 2021.

[28] Arash Vahdat, Karsten Kreis, and Jan Kautz, "Score-based generative modeling in latent space," in *NeurIPS*, 2021, pp. 11287–11302.

[29] Diederik P. Kingma and Prafulla Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *NeurIPS*, 2018, pp. 10236–10245.

[30] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski, "StyleAlign: Analysis and applications of aligned stylegan models," in *ICLR*, 2022.