PROGRESSIVE IMAGE SYNTHESIS FROM SEMANTICS TO DETAILS WITH DENOISING DIFFUSION GAN

Anonymous authors
Paper under double-blind review

ABSTRACT

Image generation has been dominated by generative adversarial Networks (GANs) due to its superior ability to generate realistic images. Recently, by decomposing the image generation process into a sequence of denoising steps, denoising diffusion probabilistic models (DDPMs) have shown remarkable sample quality and diversity in image generation. However, DDPMs typically face two main challenges (but GANs do not): the time-expensive sampling process and the semantically meaningless latent space. Although these two challenges start to draw attention in recent works on DDPMs, they are often addressed separately. In this paper, by interpreting the sampling process of DDPMs in a new way with a special noise scheduler, we propose a novel progressive training pipeline to address these two challenges simultaneously. Concretely, when the DDPMs try to predict the real images at each time step, we choose to decompose the sampling process into two stages: generating semantics firstly and then refining details progressively. As a result, we are able to interpret the sampling process of DDPMs as a refinement process instead of a denoising process. Motivated by such new interpretation, we present a novel training pipeline that progressively transforms the attention from semantics to sample quality during training. Extensive results on two benchmarks show that our proposed diffusion model achieves competitive results with as few as two sampling steps on unconditional image generation. Importantly, the latent space of our diffusion model is shown to be semantically meaningful, which can be exploited on various downstream tasks (e.g., attribute manipulation).

1 INTRODUCTION

Image generation falls in the most popular research fields in computer vision, which has been dominated by GANs in the past few years (Karras et al., 2018; 2019; 2020b; Anokhin et al., 2021) due to its superior ability to generate realistic images. Recently, Denoising Diffusion Probabilistic Models (DDPMs) (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Choi et al., 2021) have also achieved impressive results in various generation tasks, including image generation (Ho et al., 2022), audio generation (Chen et al., 2021), and 3D point cloud generation (Luo & Hu, 2021). More recent works bring further improvements to DDPMs, and show that the generation quality of DDPMs is comparable to that of GANs. In addition, DDPMs resort to likelihood computation and thus do not suffer from mode-collapse and training instability like GANs.

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. Different from GANs that synthesize images with a single forward pass through learned generator, DDPMs decompose the image generation process into a sequence of denoising steps. Therefore, DDPMs require hundreds of forward passes to generate high-quality images during inference phase, which is rather time expensive. In addition, a latent variable can be denoised to different real images by removing different noises at each step (but sampled from the same distribution) in the sampling process, resulting in that one latent variable is mapped to various real images and the latent space of DDPMs thus typically lacks high-level semantics and other desirable properties (but often possessed by GANs). Some recent works (Song et al., 2021c;a; Watson et al., 2022; Preechakul et al., 2021) start to address these two challenges, respectively. For example, Denoising Diffusion GAN (Xiao et al., 2022) is proposed to address the slow sampling by removing Gaussian assumption and adopting conditional GAN to model the denoising distribution with an expressive

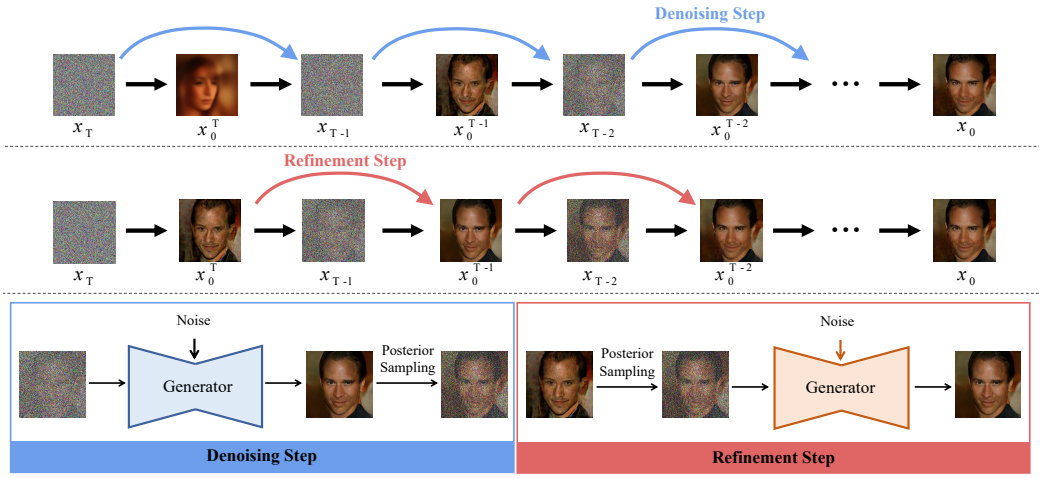

Figure 1: Two different interpretations of the sampling process of DDPMs. The top row shows the traditional interpretation as a sequence of denoising steps (with their noise scheduler). The second row shows our interpretation as a sequence of refinement steps (with our special noise scheduler). Note that these two interpretations are indeed two different partitions of the same generation process.

multimodal distribution, which takes only two sampling steps but achieves competitive sample quality and mode coverage w.r.t. traditional DDPMs. Diffusion Autoencoders (Preechakul et al., 2021) adopt an auxiliary encoder to obtain semantically meaningful representations and generate images conditioned on them. However, all of these methods can address only one of the two challenges.

In this paper, by interpreting the sampling process of DDPMs in a new way with a special noise scheduler, we propose a novel progressive training pipeline to address these two challenges simultaneously. As shown in Figure 1 (top row), the traditional DDPMs consider the sampling process as a sequence of denoising steps that progressively removes noise from the noisy images. Different from this interpretation, when DDPMs are supposed to predict the real images at each time step, the sampling process with the special noise scheduler (see Sec. 3.1) can be interpreted as a sequence of refinement steps that generates semantics firstly and then refines details progressively (see Figure 1 (second row)), resulting in that the sampling process is decomposed into the semantics generation stage and detail refinement stage. Note that these two interpretations are indeed two different partitions of the same generation process. Motivated by such interpretation, we present a novel training pipeline that progressively transforms the attention from semantics to sample quality. In particular, we introduce an auxiliary encoder to encode the input image into latent vector similar to Diffusion Autoencoders (Preechakul et al., 2021), and then enforce the generator to recover semantics from the pure Gaussian noise conditioned on the latent vector at the first step (i.e., the semantics generation stage). At the other steps, we enforce the generator to refine the details progressively while preserving the main semantic information of the output of the first step. Importantly, we choose to train the model by progressively transforming the attention from semantics generation stage to detail refinement stage. With such training pipeline, our model can achieve competitive sample quality (with as few as two sampling steps) while possessing the semantically meaningful latent space.

Our main contributions are three-fold: (1) We are the first to interpret the sampling process of DDPMs as a sequence of refinement steps that generates semantics firstly and then refines details progressively. (2) To address the two main challenges of DDPMs simultaneously, we present a novel training pipeline based on the new interpretation, which progressively transforms the attention from semantics to sample quality during training. (3) Extensive results show that our proposed diffusion model achieves competitive results with as few as two sampling steps on unconditional image generation. Importantly, the latent space of our model is shown to be semantically meaningful, which can be exploited on various downstream tasks (e.g., attribute manipulation).

2 Background

Denoising Diffusion Probabilistic Models (DDPMs) (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) belong to a family of generative models that decompose the generation process into a sequence of denoising steps. Concretely, DDPMs define a Markovian forward process that

gradually adds noise with variance β_t (at step t) to the real data $x_0 \sim q(x_0)$ in T steps:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \qquad q(x_t|x_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}). \tag{1}$$

When T is infinite and the step size β_t is infinitesimal, the reverse process (sampling process) can be described using the same functional form as the forward process (i.e., Gaussian distribution):

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \qquad p_\theta(x_{t-1}|x_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)), \tag{2}$$

where θ denotes the parameters of the denoising model, $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ and $\Sigma_\theta(\boldsymbol{x}_t, t)$ are the mean and variance of the Gaussian distribution predicted by the denoising model, and $p(x_T) = \mathcal{N}(\boldsymbol{x}_T; 0, \boldsymbol{I})$. Note that this assumption is never satisfied in practice, and thus the DDPMs can only make approximation. Generally, T is larger, the DDPMs are more accurate. Ho et al. (2020) propose to optimize the usual variational bound on negative log likelihood for training:

$$\mathcal{L} = \sum_{t \geq 1} \mathbb{E}_q[D_{KL}(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t)) + C], \tag{3}$$

where D_{KL} is the KL divergence and C denotes the constant term independent of θ. With the Gaussian assumption, this bound can be simplified as:

$$\mathcal{L}_{simple} = \sum_{t \geq 1} \mathbb{E}_{t, x_t, \epsilon}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2], \tag{4}$$

where ϵ is the noise added to x_0 to produce x_t, and ϵ_θ denotes the noise predicted by the generator. To reduce the number of sampling steps, the state-of-the-art model Denoising Diffusion GAN (Xiao et al., 2022) proposes to remove the Gaussian assumption and adopts GANs to directly approximate the true denoising distribution $q(x_{t-1}|x_t)$:

$$\min_\theta = \sum_{t \geq 1} \mathbb{E}_q[D_{adv}(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t))], \tag{5}$$

where D_{adv} is the softened reverse KL divergence (Shannon et al., 2020). Different from previous works that adopt generator to predict the added noise ϵ and derive the mean of the posterior distribution, Denoising Diffusion GAN adopts the generator to predict the x_0 firstly and then uses the posterior distribution $q(x_{t-1}|x_t, x_0)$ to sample x_{t-1}. Although this model can achieve comparative sample quality w.r.t. traditional DDPMs while takes as few as two steps for sampling, it does not address another main challenge of DDPMs (i.e., the latent space lacks high-level semantics).

3 Progressive Denoising Diffusion GAN

In this section, we firstly introduce a new interpretation of the sampling process of DDPMs with a special noise scheduler (see Sec. 3.1). Motivated by this interpretation, we decompose the sampling process into two stages: semantics generation stage and detail refinement stage (see Sec. 3.2). Based on the two-stage sampling process, we propose a novel training pipeline that progressively transforms the attention from semantics to sample quality (see Sec. 3.3). Note that we focus on the diffusion models that adopt generator to directly predict the real images x_0^t at step t in this section. However, the diffusion models that adopt generator to predict the added noise at each time step can be considered as predicting the real images indirectly and thus are easily generalized to this case.

3.1 New Interpretation of Sampling Process of DDPMs

As we have discussed in Sec. 2, diffusion models (including traditional DDPMs and Denoising Diffusion GAN (Xiao et al., 2022)) consider the sampling process as a Markovian process that progressively removes noise from noisy images, where each step can be interpreted as a denoising step. With such interpretation, the denoising model adopts generator (e.g., UNet (Ronneberger et al., 2015)) to predict the real images from noisy images firstly, and then adopts the posterior distribution $q(x_{t-1}|x_t, x_0)$ to sample the less noisy images at each denoising step (shown in the top row of Figure 1). Different from these previous works, we propose a completely new interpretation

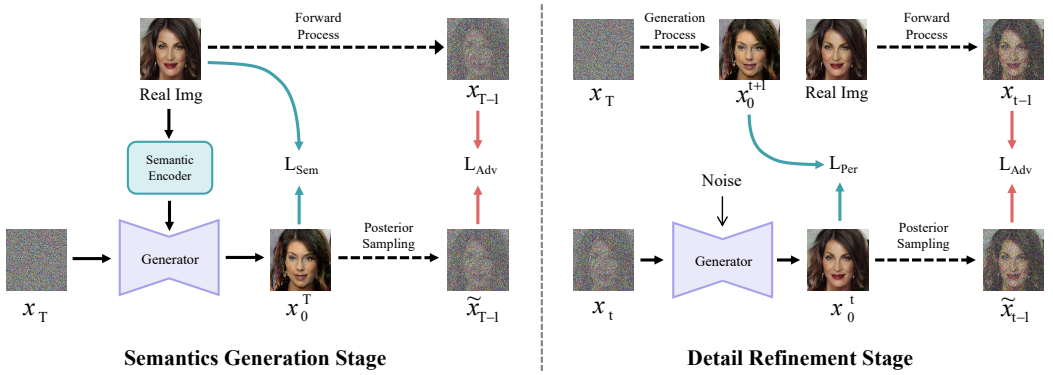

Figure 2: Illustration of the detailed training process of the semantics generation stage and the detail refinement stage. We denote the sample x_t drawn from $p_\theta(x_t|x_{t+1})$ as $\tilde{x}_t$ for easy distinction.

of sampling process in this work. Specifically, we decompose the sampling process into two stages: semantics generation stage and detail refinement stage, as shown in the second row of Figure 1. At the semantics generation stage, the generator tries to predict blur images from pure Gaussian noise, which contains main semantic information of the real images but lacks of realistic details. At the detail refinement stage, the generator gradually refines the details of the blur images generated in the semantics generation stage so that the generated images become more realistic. The refining model firstly perturbs the input image to x_{t-1} with the posterior distribution, and then predicts the more realistic image x_0^{t-1} from x_{t-1} at each refinement step. In this way, the sampling process of DDPMs can be considered as a new progressive process that generates semantics at the first step and then refines the details of generated images progressively at the left steps.

However, due to the noise scheduler of traditional DDPMs and Denoising Diffusion GAN, the semantic information of images is commonly destroyed in the middle of diffusion process (Choi et al., 2022) instead of the last step. To ensure that the semantic information is destroyed in the last step, we design a special noise scheduler $\{\beta_t|t = 1, 2, \ldots, T\}$: the diffusion step size β_t (t $< T$) is kept small so that the perturbation does not destroy the semantic information; the diffusion step size β_T is set to be large enough so that the semantic information is destroyed and the images are perturbed into pure Gaussian noise. Note that the large step size β_T does not satisfy the Gaussian assumption of traditional DDPMs. We thus remove the Gaussian assumption and adopt GAN to approximate the true denoising distribution $q(x_{t-1}|x_t)$ with small T (like Denoising Diffusion GAN). The strategy of computing such noise scheduler is given in Appendix A.

3.2 Two-stage Sampling Process

Our main goal is to address the two main drawbacks of DDPMs (i.e., the time-expensive sampling process and the semantically meaningless latent space) simultaneously. Based on the assumption of Denoising Diffusion GAN, T can be set to a small number, which would significantly speed up the sampling process. Therefore, in this work, we focus on making the latent space of Denoising Diffusion GAN more semantically meaningful. To achieve this, we present a novel two-stage framework in the following, which is motivated by the proposed interpretation in Sec. 3.1.

Semantics Generation Stage. With the new interpretation described above, the generator should learn to recover the semantic information of the input images in the semantics generation stage. To this end, we design a auxiliary encoder that learns to encode the input image x_0 into a semantically meaningful latent vector z_T, and adopt the generator to synthesize images conditioned on the information contained in the latent vector. The training process of the semantics generation stage is shown in Figure 2 (left). Formally, the input image x_0 is fed into the semantic encoder SE to derive the latent vector $z_T = SE(x_0)$. The generator G takes the noisy image x_T (derived from the forward process) as input and adopts adaptive group normalization layers (AdaGN) (Dhariwal & Nichol, 2021) to absorb the information contained in the latent vector z_T.

Detail Refinement Stage. With the outputs of the semantics generation stage x_0^T, the goal of detail refinement stage is to refine its details progressively while preserving the semantic information of it. Note that the perturbed image x_t contains the semantic information of input image in the training phase due to the small β_t (t $< T$). The semantic information in z_T is thus redundant in this stage.

Therefore, we inject the random noise $z \sim \mathcal{N}(0, \boldsymbol{I})$ to the generator G other than z_T, which is found to bring boost to the sample quality (Karras et al., 2019; Xiao et al., 2022). The training process of each step in detail refinement stage is shown in Figure 2 (right).

Learning Objective. Following Denoising Diffusion GAN (Xiao et al., 2022)), we adopt an adversarial loss to minimize the divergence $D_{KL}(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t))$:

$$\mathcal{L}_{Adv} = \sum_{t\geq 1} \mathbb{E}_{q(x_t)}\big[\mathbb{E}_{q(x_{t-1}|x_t)}[-\log(D_\phi(x_{t-1}, x_t, t))] + \mathbb{E}_{p_\theta(x_{t-1}|x_t)}[-\log(1-D_\phi(x_{t-1}, x_t, t))]\big], \quad (6)$$

where D_ϕ denotes the discriminator, and $p_\theta(x_{t-1}|x_t)$ is defined as:

$$p_\theta(x_{t-1}|x_t) = \int p_\theta(x_0|x_t)q(x_{t-1}|x_t, x_0)d\boldsymbol{x}_0 = \int p(\tilde{z})q(x_{t-1}|x_t, x_0 = G(x_t, \tilde{z}, t))d\tilde{z}, \quad (7)$$

$$p(\tilde{z}) = \begin{cases} \mathcal{N}(0, \boldsymbol{I}), & t < T \\ p(SE(x_0)), & t = T \end{cases}. \quad (8)$$

The similar idea of introducing an auxiliary encoder to learn to encode the input images into semantically meaningful latent vectors is explored in Diffusion Autoencoders (Preechakul et al., 2021), where a conditional Denoising Diffusion Implicit Model (DDIM) is designed to decode noise conditioned on the latent vectors and then the model is trained by minimizing the loss function $\|\epsilon_\theta(x_t, t, z_{sem}) - \epsilon\|_2^2$ with the Gaussian assumption. Note that Diffusion Autoencoders can learn the semantically meaningful latent space well without extra objectives due to this special loss function (L2 loss), which implicitly enforces x_t to be close to x_0 in semantics. Differently, we follow Denoising Diffusion GAN (Xiao et al., 2022) to remove the Gaussian assumption for fast sampling and thus adopt the adversarial loss for training instead of the L2 loss. Importantly, the adversarial loss only guarantees the two distribution $q(x_{t-1}|x_t)$ and $p_\theta(x_{t-1}|x_t)$ to be close, but can not guarantee the two images x_t and x_0 to be close in semantics. Therefore, to make the latent space semantically meaningful, the output of the semantics generation stage x_0^T is subject to the L1 constraints w.r.t. the input image x_0 at both pixel level and feature-map level, which encourage the semantic encoder SE to learn to extract semantically meaningful latent vector and the generator G to learn to synthesize images containing corresponding semantics conditioned on the latent vector:

$$\mathcal{L}_{Sem} = \mathbb{E}_{q(x_0)q(x_T|x_0)}\big[\|G(x_T, SE(x_0), T) - x_0\|_1 + \|V(G(x_T, SE(x_0), T)) - V(x_0)\|_1\big], \quad (9)$$

where $V(\cdot)$ denotes the function to extract feature map with pre-trained VGG network (Simonyan & Zisserman, 2014). To further guarantee the semantics of x_0^T is preserved in the detail refinement stage, we also apply the L1 constraint to the output of each refinement step:

$$\mathcal{L}_{Per} = \mathbb{E}_{q(x_0)q(x_t|x_0)p_\theta(x_{t+1:T})}\big[\|G(x_t, z, t) - x_0^{t+1}\|_1\big], \quad (10)$$

where x_0^{t+1} is sampled from $p_\theta(x_{t+1:T})$ with x_0 as input.

3.3 Progressive Training Pipeline

Although the sampling process of DDPMs is decomposed into two stages, we follow the traditional DDPMs to train our model end-to-end. Since each denoising step is independent to the other steps for DDPMs, they randomly sample t from the uniform distribution for training. Note that the generator G must learn to synthesize images that preserve the semantic information of x_0^{t+1} (see Eq. (10)), but x_0^{t+1} can not provide precise semantic information at the beginning of training due to the inaccurate sampling process $p_\theta(x_{t+1:T})$. To alleviate this issue, we present a novel progressive training pipeline, which progressively transforms the attention from semantics to sample quality. Concretely, we pick the step in the semantics generation stage (i.e., T) with the probability P and the steps in the detail refinement stage (i.e., $t < T$) with probability $1 - P$. The probability P linearly decreases as the number of training epochs increases:

$$P = P_{max} - (P_{max} - P_{min})\frac{n}{N}, \quad (11)$$

where n and N denote the current and total number of training epochs, respectively. P_{max} and P_{min} are two hyperparameters that control the trade-off between semantics and sample quality. For $t < T$, we randomly sample t from the uniform distribution, i.e., $t \sim \text{Uniform}(\{1, \dots, T-1\})$. With such dynamic sampling strategy, the model pays more attention to semantics generation at the beginning of training. As the training process goes on, the model gradually diverts attention to detail refinement. When this happens, the model has learnt to generate semantically meaningful image x_0^T, and the refinement steps thus can be trained efficiently based on the semantic information of x_0^T.

4 Related Works

Denoising Diffusion Probabilistic Models (DDPMs) (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) aim to learn to generate data from the pure Gaussian noise by the finite-time reversal of a diffusion process, which transits noisy data to less noisy data in each step. Due to its superior performance on sample quality and mode coverage, DDPMs have been deployed in various generation tasks, such as unconditional image generation (Ho et al., 2020; Dhariwal & Nichol, 2021; Nichol & Dhariwal, 2021), text-to-image generation (Nichol et al., 2022; Rombach et al., 2021), image super-resolution (Saharia et al., 2021; Li et al., 2022), and text-driven image editing (Avrahami et al., 2021). However, DDPMs typically face two challenges: the sampling process of DDPMs is time-expensive, and the latent space of DDPMs lacks high-level semantics.

To address the first challenge, a number of recent methods have been proposed. Song et al. (2021a); Watson et al. (2022); Kong & Ping (2021); Jolicoeur-Martineau et al. (2021a) employ various sampling strategies to reduce the number of sampling steps, but still require hundreds of sampling steps to generate high-quality images. Luhman & Luhman (2021) adopt knowledge distillation to convert DDPM into a new model that can generate images in one step, resulting in inferior sample quality w.r.t. GANs. Xiao et al. (2022) balance the efficiency of sampling process, the sample quality and the mode coverage by removing the Gaussian assumption and modeling the denoising distribution with an expressive multimodal distribution, which reduces the required steps in both training and inference phases while achieving competitive results in sample quality and mode coverage.

However, these methods focus on addressing the first challenge, and ignore the semantics of the latent space of DDPMs. Note that semantically meaningful latent space is widely explored in other deep generative models (e.g., GAN (Goodfellow et al., 2020) and VAE (Kingma & Welling, 2013)), which is useful for various downstream tasks (e.g., controllable image manipulation with GAN (Patashnik et al., 2021; Yang et al., 2021; Zhu et al., 2020) and model-based reinforcement learning with VAE (Hafner et al., 2019; Freeman et al., 2019; Ha & Schmidhuber, 2018)). To make the latent variables more meaningful, Diffusion Autoencoders (Preechakul et al., 2021) design an auxiliary encoder to encode the input image into a semantically meaningful latent vector and adopt DDIM (Song et al., 2021a) to reconstruct the input image conditioned on the latent vector, which requires hundreds of sampling steps to generate high-quality images. In this paper, by interpreting the sampling process of DDPMs in a completely new way with the special noise scheduler, we present a novel progressive training pipeline to address the two main challenges of DDPMs simultaneously.

5 Experiments

5.1 Datasets and Settings

Datasets. To evaluate the effectiveness of our proposed diffusion model, we mainly conduct experiments on the CelebA-HQ (Karras et al., 2018) dataset, which has 30,000 high-quality face images of the resolution 256×256. Moreover, we also conduct experiments on the CIFAR-10 (Krizhevsky, 2009) dataset, since it is widely used in previous works. CIFAR-10 consists of 60,000 diverse images from 10 different classes. We follow the standard training/test split of each dataset.

Implementation Details. Our proposed diffusion model is implemented on top of Denoising Diffusion GAN (Xiao et al., 2022), which adopts UNet as the generator and a time-independent CNN as the discriminator. Moreover, the number of time steps T is set to 2 in all experiments. P_{max} and P_{min} are empirically set to 0.8 and 0.4, respectively. We use 8 V100 GPUs to train our model on both CelebA-HQ and CIFAR-10, which takes about 160 hours and 50 hours, respectively. For unconditional generation, we following Diffusion Autoencoders (Preechakul et al., 2021) to train an extra diffusion model to sample the latent vectors. The code and models will be released soon.

Evaluation Metrics. We adopt the Frechét Inception Distance (FID) (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016) to evaluate the quality of generated images, which are commonly used in previous works. Following Denoising Diffusion GAN, we utilize the improved recall score (Kynkäänniemi et al., 2019) to evaluate the diversity of generated images. For fair comparison, we generate 30,000 images for CelebA-HQ and 50,000 images for CIFRA-10 during evaluation, and compute the three metrics with the generated images and images in the training set. To evaluate the efficiency of sampling process, we also report the clock time of generating a batch of 100 images on a V100 GPU and the number of function evaluations (NFE) as metrics. Note that the NFE of our model does not contain the number of steps to sample the latent vectors.

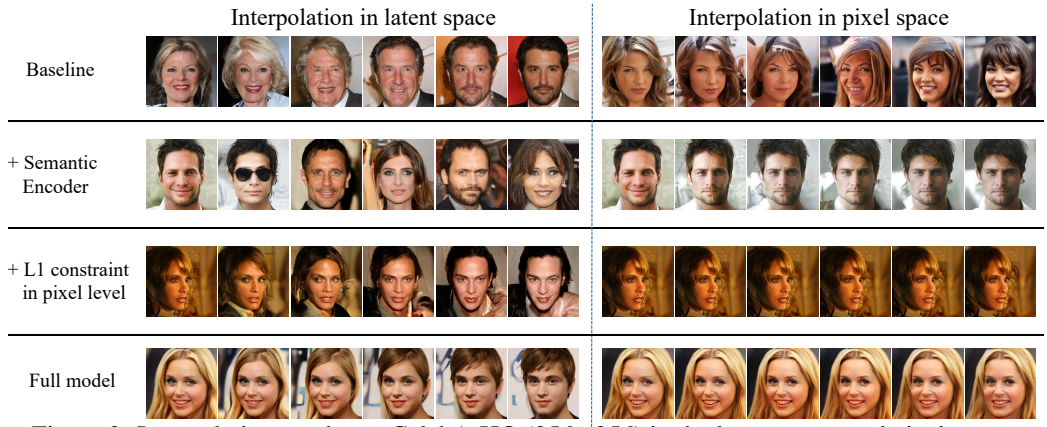

Figure 3: Interpolation results on CelebA-HQ (256×256) in the latent space and pixel space.

Table 1: The results of the ablation study of our proposed full diffusion model for unconditional generation 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.

Method	Semantic	FID ↓
Baseline	N	7.64
+ Semantic encoder	N	**5.88**
+ L1 constraint at pixel level	Y	6.77
+ L1 constraint at feature-map level	Y	6.66
+ Progressive training pipeline (ours)	Y	<u>6.47</u>

Table 2: Quantitative results for unconditional high-quality generation on CelebA-HQ.

Method	Semantic	FID ↓
NVAE (Vahdat & Kautz, 2020)	-	29.7
VAEBM (Xiao et al., 2021)	-	20.4
NCP-VAE (Aneja et al., 2021)	-	24.8
PGGAN (Karras et al., 2018)	Y	8.03
VQ-GAN (Esser et al., 2021)	Y	10.2
DC-AE (Parmar et al., 2021)	Y	15.8
Score SDE (Song et al., 2021c)	N	7.23
LSGM (Vahdat et al., 2021)	N	7.22
UDM (Kim et al., 2021)	N	7.16
LDM, T=500 (Rombach et al., 2021)	N	**5.11**
Denoising Diffusion GAN (Xiao et al., 2022)	N	7.64
P2, T=500 (Choi et al., 2022)	N	6.91
Ours, T=2	Y	<u>6.47</u>

5.2 Ablation Studies

To demonstrate the contributions of our proposed components and provide insightful analysis of our model, we conduct ablation studies on CelebA-HQ. Concretely, we consider the Denoising Diffusion GAN (Xiao et al., 2022) 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 to the generator for each step t instead of the random noises. Note that the resultant model can be considered as the simple combination of Denoising Diffusion GAN and Diffusion Autoencoders (Preechakul et al., 2021). 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 pre-trained VGG network when applying the two L1 constraints to the step T only and injecting the random noise for $t < T$ (denoted as '+ L1 constraint at feature map level' for short). Finally, we add the constraint in Eq. (10) to the model and train it with the proposed progressive training pipeline (denoted as '+ progressive training pipeline' for short).

Semantics. To the best of our knowledge, there is no effective metric to evaluate how the space is semantically meaningful. Therefore, we indirectly evaluate this by exploring the interpolation results of ablated models (i.e., the interpolation results are more smooth, the latent space is more semantically meaningful), which is commonly used in the literature (Preechakul et al., 2021; Wu et al., 2022; Kingma & Dhariwal, 2018). For baseline and its variants, both the latent space and pixel space could contain high-level semantics, and thus we explore these two spaces separately. Note that the variables in the latent space of baseline are the random noises injected at each step, while that of the variants are the vectors extracted by the semantic encoder. Concretely, we randomly generate two images with two random noises in the latent space (or pixel space), while inputting the same noises in the other space. We then generate images with the linear interpolation of the two random noises. The interpolation results are shown in Figure 3. We can observe that: (1) The interpolation results of the baseline are not smooth in both latent space and pixel space, indicating that the two spaces of the baseline are semantically meaningless (first row). (2) Adding the semantic encoder on the top of baseline makes the main semantics of generated images be controlled by the variables in the latent space. However, the latent space is still semantically meaningless. Although the results

Table 3: Quantitative results for unconditional generation on CIFAR-10.

Method	Semantic	IS ↑	FID ↓	Recall ↑	NFE ↓	Time (s) ↓
Glow (Kingma & Dhariwal, 2018)	Y	3.92	48.9	-	1	-
PixelCNN (van den Oord et al., 2016)	N	4.60	65.9	-	1024	-
NVAE (Vahdat & Kautz, 2020)	-	7.18	23.5	0.51	1	0.36
IGEBM (Du & Mordatch, 2019)	Y	6.02	40.6	-	60	-
VAEBM (Xiao et al., 2021)	-	8.43	12.2	0.53	16	8.79
SNGAN (Miyato et al., 2018)	-	8.22	21.7	0.44	1	-
SNGAN+DGflow (Ansari et al., 2021)	-	9.35	9.62	0.48	25	1.98
TransGAN (Jiang et al., 2021)	Y	9.02	9.26	-	1	-
StyleGAN2 w/o ADA (Karras et al., 2020a)	Y	9.18	8.32	0.41	1	0.04
StyleGAN2 w/ ADA (Karras et al., 2020a)	Y	9.83	2.92	0.49	1	0.04
StyleGAN2 w/ Diffaug (Zhao et al., 2020)	Y	9.40	5.79	0.42	1	0.04
DDPM (Ho et al., 2020)	N	9.46	3.21	0.57	1000	80.5
NCSN (Song & Ermon, 2019)	N	8.87	25.3	-	1000	107.9
Adversarial DSM (Jolicoeur-Martineau et al., 2021b)	N	-	6.10	-	1000	-
Likelihood SDE (Song et al., 2021b)	N	-	2.87	-	-	-
Score SDE (VE) (Song et al., 2021c)	N	9.89	2.20	0.59	2000	423.2
Score SDE (VP) (Song et al., 2021c)	N	9.68	2.41	0.59	2000	421.5
Probability Flow (VP) (Song et al., 2021c)	N	9.83	3.08	0.57	140	50.9
LSGM (Vahdat & Kautz, 2020)	N	9.87	**2.10**	0.61	147	44.5
DDIM, T=50 (Song et al., 2021a)	N	8.78	4.67	0.53	50	4.01
FastDDPM, T=50 (Kong & Ping, 2021)	N	8.98	3.41	0.56	50	4.01
Recovery EBM (Gao et al., 2021)	N	8.30	9.58	-	180	-
Improved DDPM (Nichol & Dhariwal, 2021)	N	-	2.90	-	4000	-
VDM (Kingma et al., 2021)	N	-	2.90	-	4000	-
UDM (Kim et al., 2021)	N	**10.1**	2.33	-	2000	-
D3PMs (Austin et al., 2021)	N	8.56	7.34	-	1000	-
Gotta Go Fast (Jolicoeur-Martineau et al., 2021a)	N	-	4.00	-	1000	-
DDPM Distillation (Luhman & Luhman, 2021)	N	8.36	9.36	0.51	1	-
Analytic-DDPM (Bao et al., 2022)	N	-	4.11	-	10	-
DPM-solver (Lu et al., 2022)	N	-	5.28	-	12	-
Denoising Diffusion GAN (Xiao et al., 2022)	N	9.63	3.75	0.57	4	0.21
Ours	Y	9.48	4.08	**0.62**	2	0.16

in the pixel space are more smooth, it only affects minor semantics (second row). (3) Adding the L1 constraint at pixel level further makes the semantics of generated images be controlled by the variables in the latent space, while the variables in the pixel space only affects minor details (semantically meaningless). Notice that the interpolation results in the latent space are smooth, i.e., it is semantically meaningful (third row). This finding holds for our full model (fourth row).

Sample Quality. We explore how our proposed components contribute to the sample quality in unconditional generation. The results of ablation study are shown in Table 1. It can be seen that: (1) The simple combination of Denoising Diffusion GAN and Diffusion Autoencoder can bring a boost in FID. However, it can not make the latent space semantically meaningful (see Figure 3). (2) Adding the L1 constraint at pixel level makes the latent space semantically meaningful at the cost of sample quality, indicating that forcing the model to reconstruct the input image with the latent vector extracted by the auxiliary encoder is essential to learn semantically meaningful latent space for DDPMs. (3) Our simple but effective modification brings a boost in FID. Importantly, our proposed progressive training pipeline can further improve the FID (but ablating the constraint in Eq. (10) from this progressive pipeline leads to a drop in our extra experiments).

5.3 Comparison to the State-of-the-Arts

Furthermore, we compare our diffusion model with the state-of-the-arts in the unconditional generation tasks. The quantitative results on CelebA-HQ (that contains high-resolution images) are shown in Table 2. We can observe that our diffusion model outperforms all the competitors except LDM (Rombach et al., 2021) in unconditional high-resolution generation. However, LDM requires 500 steps in sampling process, which takes about $\mathbf{210}$ seconds to generate a batch of 100 images on a V100 GPU (our model takes only $\mathbf{2.3}$ seconds). Furthermore, we also make comparison to the state-of-the-art methods on CIFAR-10, and the results are shown in Table 3. We can observe that our diffusion model achieves competitive results in sample quality (IS and FID) and mode coverage (Recall) but with much fewer sampling steps. Note that all of the DDPM-based methods that outperform our diffusion model in sample quality require more than 50 steps for sampling (except Denoising

Figure 4: Samples generated by our model on CelebA-HQ (256×256) and CIFAR-10 (32×32).

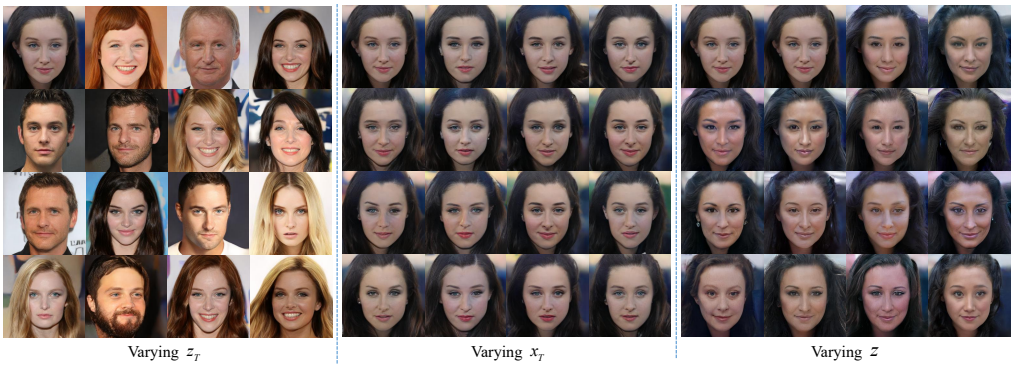

Figure 5: Samples generated by varying one of the three inputs of our diffusion model on CelebA-HQ (256×256) while keeping the other two inputs unchanged. Note that the top-left images of the three columns/groups are exactly the same due to the same inputs.

Diffusion GAN), but our diffusion model requires only two sampling steps to achieve competitive results, which speeds up the sampling process more than 20×. In particular, our diffusion model achieves competitive sample quality w.r.t. Denoising Diffusion GAN (our baseline model) with fewer steps and outperforms it in mode coverage. Importantly, the sample quality of our model outperforms that of Denoising Diffusion GAN on CelebA-HQ, indicating that our proposed method is more effective on generating high-resolution images. We show the qualitative results on CelebA-HQ and CIFAR-10 in Figure 4. More generation samples are given in Appendix B.

5.4 Additional Studies

Finally, we explore how the three inputs of our model (i.e., z_T, z and x_T) affect the content of images generated by our model. Concretely, we generate images by varying only one of the three inputs each time while keeping the other two inputs unchanged. The generated samples are shown in Figure 5. We can observe that the semantics of generated images is mainly controlled by z_T. With the constraints in Eq. (9) and Eq. (10), x_T can only affect minor details of the generated image. The random noise z affects some obvious details of the generated images (e.g., the color of skin and eyes), but does not affect the main semantics of generated images. Note that this issue can be alleviated by injecting z_T at each time step, which leads to worse sample quality.

6 Conclusion

In this work, we firstly introduce a new interpretation of sampling process of DDMPs. Based on this interpretation, 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 on two benchmark show that our proposed model can achieve comparative results with as few as two sampling steps on unconditional images generation. Importantly, our diffusion model can generate smooth interpolation results, indicating that the latent space of our model is semantically meaningful and it is readily exploited on various downstream tasks.

REFERENCES

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

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

Abdul Fatir Ansari, Ming Liang Ang, and Harold Soh. Refining deep generative models via discriminator gradient flow. In *ICLR*, 2021.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, pp. 17981–17993, 2021.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. *arXiv preprint arXiv:2111.14818*, 2021.

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *ICLR*, 2022.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating gradients for waveform generation. In *ICLR*, 2021.

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

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.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, pp. 8780–8794, 2021.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *NeurIPS*, pp. 3603–3613, 2019.

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

C. Daniel Freeman, David Ha, and Luke Metz. Learning to predict without looking ahead: World models without forward prediction. In *NeurIPS*, pp. 5380–5391, 2019.

Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P. Kingma. Learning energy-based models by diffusion recovery likelihood. In *ICLR*, 2021.

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

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, pp. 2455–2467, 2018.

Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, volume 97, pp. 2555–2565, 2019.

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*, pp. 6626–6637, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pp. 6840–6851, 2020.

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*, 23:47:1–47:33, 2022.

Yifan Jiang, Shiyu Chang, and Zhangyang Wang. TransGAN: Two transformers can make one strong GAN. *arXiv preprint arXiv:2102.07074*, 2021.

Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021a.

Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Ioannis Mitliagkas, and Remi Tachet des Combes. Adversarial score matching and improved sampling for image generation. In *ICLR*, 2021b.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

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

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, pp. 8107–8116, 2020b.

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.

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

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

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *arXiv preprint arXiv:2107.00630*, 2021.

Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, pp. 3929–3938, 2019.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. SRDiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, pp. 2837–2845, 2021.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, volume 139, pp. 8162–8171, 2021.

Alexander Quinn 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. In *ICML*, volume 162, pp. 16784–16804, 2022.

Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. Dual contradistinctive generative autoencoder. In *CVPR*, pp. 823–832, 2021.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *ICCV*, pp. 2065–2074, 2021.

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

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241, 2015.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, pp. 2226–2234, 2016.

Matt Shannon, Ben Poole, Soroosh Mariooryad, Tom Bagby, Eric Battenberg, David Kao, Daisy Stanton, and R. J. Skerry-Ryan. Non-saturating GAN training as divergence minimization. *arXiv preprint arXiv:2010.08029*, 2020.

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

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, volume 37, pp. 2256–2265, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, pp. 11895–11907, 2019.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *NeurIPS*, pp. 1415–1428, 2021b.

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*, 2021c.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020.

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

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, volume 48, pp. 1747–1756, 2016.

Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. Learning fast samplers for diffusion models by differentiating through sample quality. In *ICLR*, 2022.

Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. StyleAlign: Analysis and applications of aligned stylegan models. In *ICLR*, 2022.

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

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *ICLR*, 2022.

Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. L2M-GAN: learning to manipulate latent space semantics for facial attribute editing. In *CVPR*, pp. 2951–2960, 2021.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. In *NeurIPS*, 2020.

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain GAN inversion for real image editing. In *ECCV*, volume 12362, pp. 592–608, 2020.

A Strategy of Computing β_t

The proposed novel interpolation requires a special noise scheduler $\{\beta_t | t = 1, 2, \ldots, T\}$ so that the diffusion model can recover the semantics at the T step. In this section, we introduce a simple strategy to compute such special noise scheduler.

The main idea of our strategy is to compress all steps $t > \hat{t}$ in the common scheduler into one step, where $\hat{t}$ denotes the step that destroys the semantics. Concretely, following Denoising Diffusion GAN (Xiao et al., 2022), we define the initial noise scheduler as $\{\beta_t | t = 1, 2, \ldots, \hat{T}\}$ ($\hat{T} > T$), which is computed based on the continuous-time diffusion model formulation:

$$\beta_t = 1 - \boldsymbol{e}^{-\beta_{min}(\frac{1}{\hat{T}}) - 0.5(\beta_{max} - \beta_{min})\frac{2t-1}{\hat{T}^2}}, \tag{12}$$

where β_{max} and β_{min} are set to 20 and 0.1, respectively. The value of $\hat{t}$ is derived by the value signal-to-noise ratio (SNR). As stated in Choi et al. (2022), the model leanrs the content when SNR is between 10^{-2} and 10^{0}. We thus empirically set the step $\hat{t}$ as that SNR($\hat{t}$) $> 10^{-1}$ and SNR($\hat{t}+1$) $< 10^{-1}$. To ensure that the images are destroyed at the step $\hat{t}$, we set the $\beta_{\hat{t}}$ as:

$$\beta_{\hat{t}} = 1 - \frac{\bar{\alpha}_{\hat{T}}}{\bar{\alpha}_{\hat{t}-1}}. \tag{13}$$

We then discard the steps $t > \hat{t}$, and set the variance of noise of $t < \hat{t}$ to β_t, which is the same as that in the initial noise scheduler. The formed special noise scheduler is formally defined as $\{\beta_1, \beta_2, \ldots, \beta_{\hat{t}}\}$. Note that we can not determine the value of $\hat{t}$ before computing the SNR. Therefore, to obtain the special noise scheduler for desired T with this strategy, we have to compute the SNR for multiple noise schedulers $\{\beta_t | t = 1, 2, \ldots, \hat{T}\}$ with different $\hat{T}$ so that we can determine $\hat{t} = T$. Other strategies to compute such special noise scheduler are possible, but we do not explore them because the strategy to compute β_t is not the scope of this paper.

B Additional Quality Results

We show additional samples of unconditional generation on the CelebA-HQ dataset in Figure 6. We also give additional interpolation results of our full model in Figure 7, indicating that the latent space of our model indeed contains high-level semantics.

Figure 6: Additional samples generated by our model on CelebA-HQ (256×256).

Figure 7: Additional interpolation results of our full model. Each row shows the interpolation between the the leftmost image and the rightmost image.