Supplementary of "Semi-Implicit Denoising Diffusion Models (SIDDMs)"

Anonymous Author(s)
Affiliation
Address
email

S I. Derivation of Training Objective

Before getting the final training objective, we formulate the forward posterior following [1, 2]. Via Bayes' rule, we can rewrite the forward posterior given $\mathbf{x}_0$:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)},$$

where all the components of the very right equation are forward diffusion and follow Gaussian distribution. Thus the forward posterior can be rewritten as Gaussian with mean and standard deviation as follows:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1};\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0),\tilde{\beta}_t\mathbf{I}).$$

Here we do not give redundant derivation and give the firm of the forward posterior given x_t, x_0.

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t.$$

And we parameterize our denoised x'_{t-1} given the predicted x'_0 and the input x_t via simply replacing the x_0 with the predicted x'_0 in the above equation. Then, given the parameterized x'_{t-1}.

To get our final training objective, We can rewrite the distribution matching objective of Equation (7) as:

$$\begin{aligned}
&\min_\theta \max_{D_{adv}} \mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}\Big[D_{adv}(q(x_{t-1})||p_\theta(x_{t-1})) \\
&\qquad\qquad + \lambda_{AFD}[-H(p_\theta(x_t|x_{t-1})) + H(p_\theta(x_t|x_{t-1}), q(x_t|x_{t-1}))]\Big] \\
&= \min_\theta \max_{D_{adv},\psi} \mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}\Big[D_{adv}(q(x_{t-1})||p_\theta(x_{t-1})) \\
&\qquad\qquad + \lambda_{AFD}[H(p_\theta(x_t|x_{t-1}), q(x_t|x_{t-1})) - H(p_\theta(x_t|x_{t-1}), p_\psi(x_t|x_{t-1}))]\Big],
\end{aligned}$$

where the first GAN matching objective can be written as:

$$\min_\theta \max_{D_\phi} \sum_{t>0} \mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}[-\log(D_\phi(x_{t-1},t))] + [-\log(1-D_\phi(x'_{t-1},t))].$$

In the first cross-entropy of our distribution matching objective, the $q(x_t|x_{t-1})$ is the forward diffusion with the mean $\sqrt{1-\beta_t}x_{t-1}$ and variance $\beta_t\mathbf{I}$. Thus the likelihood can be written as:

$$H(p_\theta(x_t|x_{t-1}), q(x_t|x_{t-1})) = \mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}\frac{(1-\beta_t)\left\|x'_{t-1}-x_{t-1}\right\|^2}{\beta_t},$$

To solve the second cross-entropy between the denoised distribution and the parameterized regression model, we define $p_\psi(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$ as forward diffusion for the regression model. And we also define $x't$ are sampled from the x'_{t-1} via the forward diffusion. Similar to the above likelihood of cross-entropy, we can write the following likelihood for the second cross-entropy as follows:

$$H(p_\theta(x_t|x_{t-1}), p_\psi(x_t|x_{t-1})) = \mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})} \frac{\left\|C_\psi(x'_{t-1}) - x'_t\right\|^2}{\beta_t},$$

Finally, we can get the final training objective of our proposed method.

$$\min_\theta \max_{D_\phi, C_\psi} \sum_{t>0} \mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})} \Big[[-\log(D_\phi(x_{t-1}, t))] + [-\log(1 - D_\phi(x'_{t-1}, t))] + \lambda_{AFD} \frac{(1-\beta_t)\left\|x'_{t-1} - x_{t-1}\right\|^2 - \left\|C_\psi(x'_{t-1}) - x'_t\right\|^2}{\beta_t}\Big],$$

In the main paper formulation, we mistakenly exchange the position of the β_t and $1-\beta_t$, it is a typo, we will correct it later.

S II. Derivation of Theorem 1

For simplicity, we denote q via Q, p_θ via P and x_{t-1}, x_t via X, Y. According to the triangle inequality of total variation (TV) distance, we have

$$d_{TV}(Q_{XY}, P_{XY}) \leq d_{TV}(Q_{XY}, Q_{Y|X}P_X) + d_{TV}(Q_{Y|X}P_X, P_{XY}). \tag{E11}$$

Using the definition of TV distance, we have

$$\begin{aligned} d_{TV}(Q_{Y|X}Q_X, Q_{Y|X}P_X) &= \frac{1}{2}\int |Q_{Y|X}(y|x)Q_X(x) - Q_{Y|X}(y|x)P_X(x)|\mu(x,y) \\ &\overset{(a)}{\leq} \frac{1}{2}\int |Q_{Y|X}(y|x)|\mu(x,y) \int |Q_X(x) - P_X(x)|\mu(x) \\ &\leq c_1 d_{TV}(Q_X, P_X), \end{aligned} \tag{E12}$$

where P and Q are densities, μ is a (σ-finite) measure, c_1 is an upper bound of $\frac{1}{2}\int |Q_{Y|X}(y|x)|\mu(x,y)$, and (a) follows from the Hölder inequality.

Similarly, we have

$$d_{TV}(Q_{Y|X}P_X, P_{Y|X}P_X) \leq c_2 d_{TV}(Q_{Y|X}, P_{Y|X}), \tag{E13}$$

where c_2 is an upper bound of $\frac{1}{2}\int |P_X(x)|\mu(x)$. Combining (E11), (E12), and (E13), we have

$$d_{TV}(Q_{XY}, P_{XY}) \leq c_1 d_{TV}(Q_X, P_X) + c_2 d_{TV}(Q_{Y|X}, P_{Y|X}) \tag{E14}$$

According to he Pinsker inequality $d_{TV}(P,Q) \leq \sqrt{\frac{KL(P||Q)}{2}}$ [3], and the relation between TV and JSD, *i.e.*, $\frac{1}{2}d_{TV}(P,Q)^2 \leq JSD(P,Q) \leq 2d_{TV}(P,Q)$ [4], we can rewrite (E14) as

$$JSD(Q_{XY}, P_{XY}) \leq 2c_1\sqrt{2JSD(Q_X, P_X)} + 2c_2\sqrt{2KL(P_{Y|X}||Q_{Y|X})}. \tag{E15}$$

S III. Societal impact

With the increasing utilization of generative models, our proposed SSIDMs will improve the diffusion-based generative model while maintaining the highest level of generative quality. The incorporation of SSIDMs enhances the capabilities of generative models, particularly in the domain of text-to-image generation and editing. By integrating SSIDMs into the existing generative model framework, we could unlock new possibilities for generating realistic and visually coherent images from textual descriptions. One of the key advantages of our SSIDMs is their ability to accelerate the inference process, even though our model takes more time and more resources to train because of the additional adversarial training objectives. With faster inference, we eliminate the time-consuming barriers previously associated with text-to-image generation. As a result, real-time applications of generative models become feasible, enabling on-the-fly image generation or instant editing.

S IV. More Implementation Details

For the time steps, we apply the continuous time setup with the cosine noise schedule for all the experiments. We also apply a similar network structure as [5] and the downsampling trick as [6], where we put the downsampling layer at the beginning of each ResBlock. As mentioned, we design the discriminator as UNet, which adopts the symmetric network structure as the generator. For the regression model C_ψ and the discriminator regularizer, we share most of the layers with the discriminator except that we put a different linear head for the marginal, conditional and regularizer outputs. To be notified, the C_ψ only works on the denoised data, and the regularizer only works on the sampled x_{t-1} from the real data via forward diffusion. By this design, our model does not bring obvious extra overhead than our baseline DDGANs, which only has two more linear head in the final output for the discriminator network. We also describe the detailed model hyperparameters in the following table. We train all the models until they converges to the best FID score.

	CIFAR10 32	CelebA-HQ 256	ImageNet 64
Resolution	32	256	64
Conditional on labels	False	False	True
Diffusion steps	4	2	4
Noise Schedule	cosine	cosine	cosine
Channels	256	192	256
Depth	2	2	2
Channels multiple	2,2,2	1,1,2,3,4	1,2,3,4
Heads	4	4	4
Heads Channels	64	64	64
Attention resolution	16,8	32,16,8	32,16,8
Dropout	0.1	0.1	0.1
Batch size	256	128	2048
Learning Rate of G	2e-4	2e-4	2e-4
Learning Rate of D	1e-4	1e-4	1e-4
EMA Rate of G	0.9999	0.9999	0.9999

Table 1: Hyperparameters for our SSIDMs on different datasets.

S V. More Generated Results

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[2] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations*, 2022.

[3] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[4] Kiran K Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. In *NeurIPS*, pages 10271–10282. Curran Associates, Inc., 2018.

[5] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[6] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Figure 1: Randomly generated samples from our model. We randomly sample 768 images from the generated images from CIFAR10, which we used to produce our paper results

Figure 2: Randomly generated samples from our model. We randomly sample 192 images from the generated images from CeleHQ 256, which we used to produce our paper results

Figure 3: Randomly generated samples from our model. We randomly sample 768 images from the generated images from Imagenet 64, which we used to produce our paper results