Supplementary Materials of "Entropy-driven Sampling and Training Scheme for Conditional Diffusion Generation"

Guangcong Zheng[1⋆], Shengming Li[1⋆], Hui Wang[1], Taiping Yao[2], Yang Chen[2], Shouhong Ding[2], and Xi Li[1,3,4⋆⋆]

[1] College of Computer Science & Technology, Zhejiang University
{guangcongzheng, shengming22, wanghui_17, xilizju}@zju.edu.cn
[2] Youtu Lab, Tencent, China
{taipingyao, wizyangchen, ericshding}@tencent.com
[3] Shanghai Institute for Advanced Study, Zhejiang University
[4] Shanghai AI Laboratory

1 Architecture of Model

The architecture setting is presented in Implementation Details of Sec. 4.1 of main paper, where we follow all the setting as [1]. The generator of a diffusion model, usually based on an encoder-decoder architecture such as U-Net, learns to produce a slightly more "denoised" x_{t-1} from a noisy image x_t. And the classifier is just the encoder part of U-Net appended with a classification head. The U-Net in our paper is introduced by [3] and improved by [6,4,1], from which more details can be found.

2 Comparisons with SOTA Results

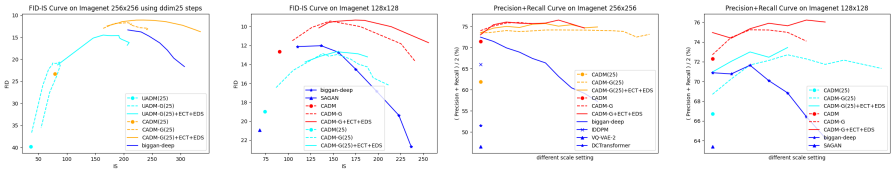


Fig. 1: The FID & IS curve and the Precision+Recall curve.

⋆ The first two authors contributed equally to this paper.
⋆⋆ The corresponding author is Xi Li.

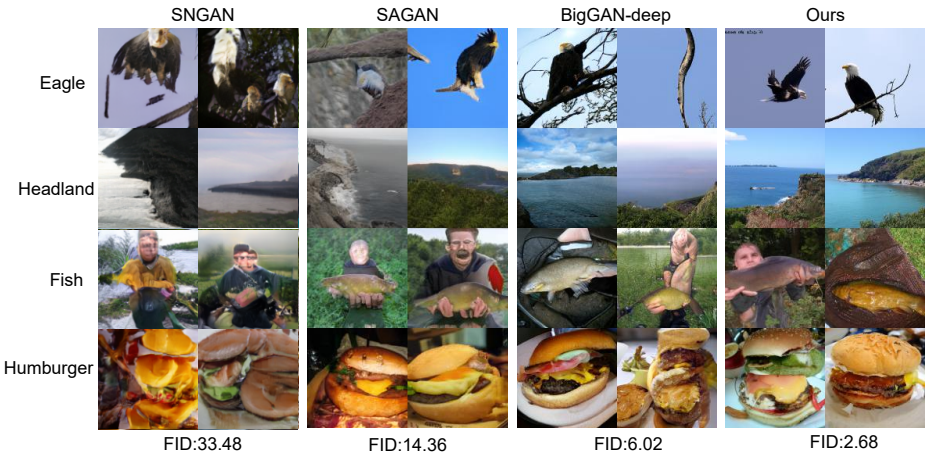


Fig. 2: Qualitatively comparision with other class-conditional methods on ImageNet 128x128.

3 More results on ImageNet1000 at 64x64, 128x128

In this section, we show the experiment results on ImageNet1000 at 64×64, 128×128 resolutions, to further verify the effectiveness of our proposed methods. From Table 1, it can be concluded that our proposed methods can adapt to low resolution image generation. All results based on previous methods in this table are cited from Dhariwal *et al.*[1], except for CADM-G (25) in ImageNet 128×128, of which the evaluation metric is achieved by our reproducing result.

Table 1: **Comparison results with state-of-the-art generative models based on ImageNet 128×128, 64×64.** Annotation '(25)' after the method means its sampling process is based on DDIM with 25 steps. Otherwise, it means the normal sampling method in DDPM, with 250 steps.

Method	FID ↓	sFID ↓	Prec ↑	Rec ↑
ImageNet 64×64				
BigGAN-deep [2]	4.06	3.96	**0.79**	0.48
IDDPM [5]	2.92	**3.79**	0.74	0.62
CADM-G [1]	2.07	4.29	0.74	**0.63**
CADM-G+EDS+ECT	**1.88**	4.51	0.76	0.61
ImageNet 128×128				
BigGAN-deep [2]	6.02	7.18	**0.86**	0.35
LOGAN [7]	3.36			
CADM [1]	5.91	5.09	0.70	**0.65**
CADM-G (25) [1]	6.58	7.52	0.77	0.50
CADM-G+EDS+ECT (25)	6.22	7.10	0.78	0.49
CADM-G [1]	2.97	**5.09**	0.78	0.59
CADM-G+EDS+ECT	**2.68**	5.10	0.80	0.56

4 More Visualizations

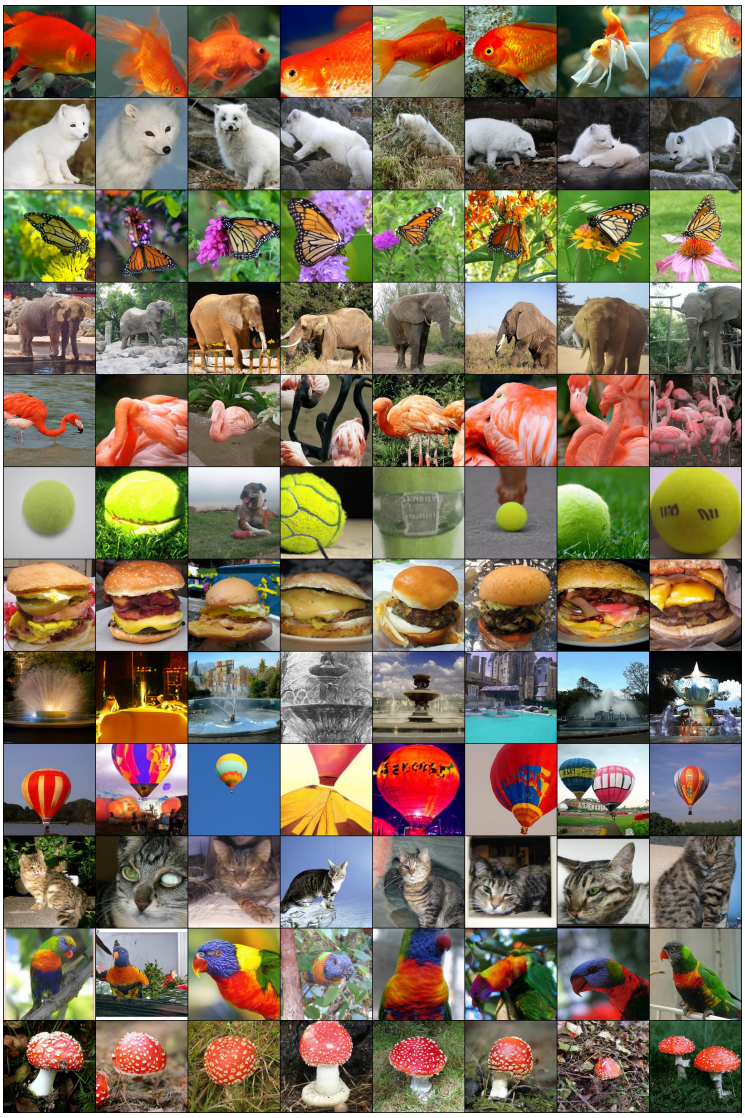


Fig. 3: Generated images in ImageNet 256×256 from CADM with EDS and ECT (DDPM 250 steps).

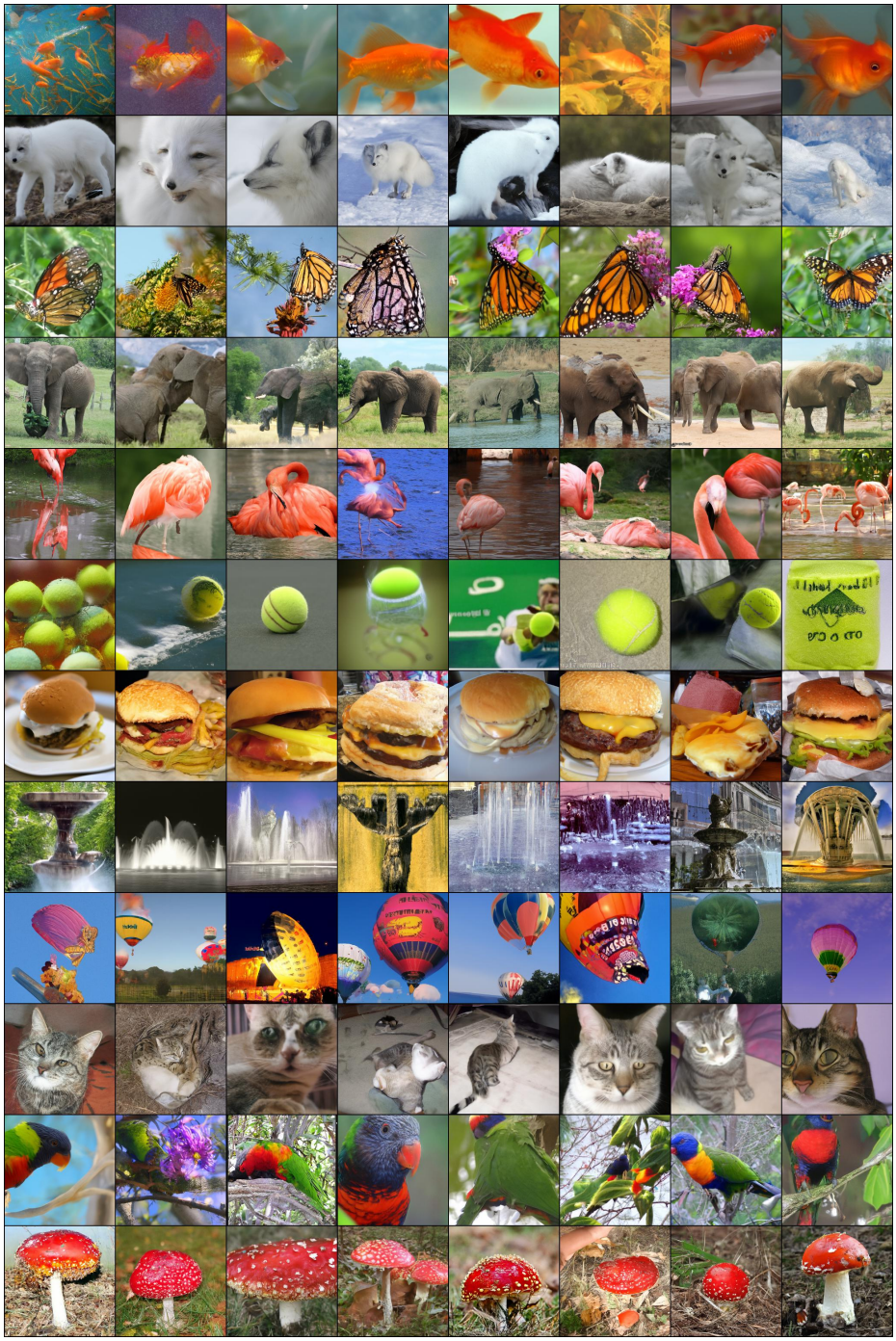


Fig. 4: Generated images in ImageNet 256×256 from CADM-G with EDS and ECT (DDIM 25 steps).

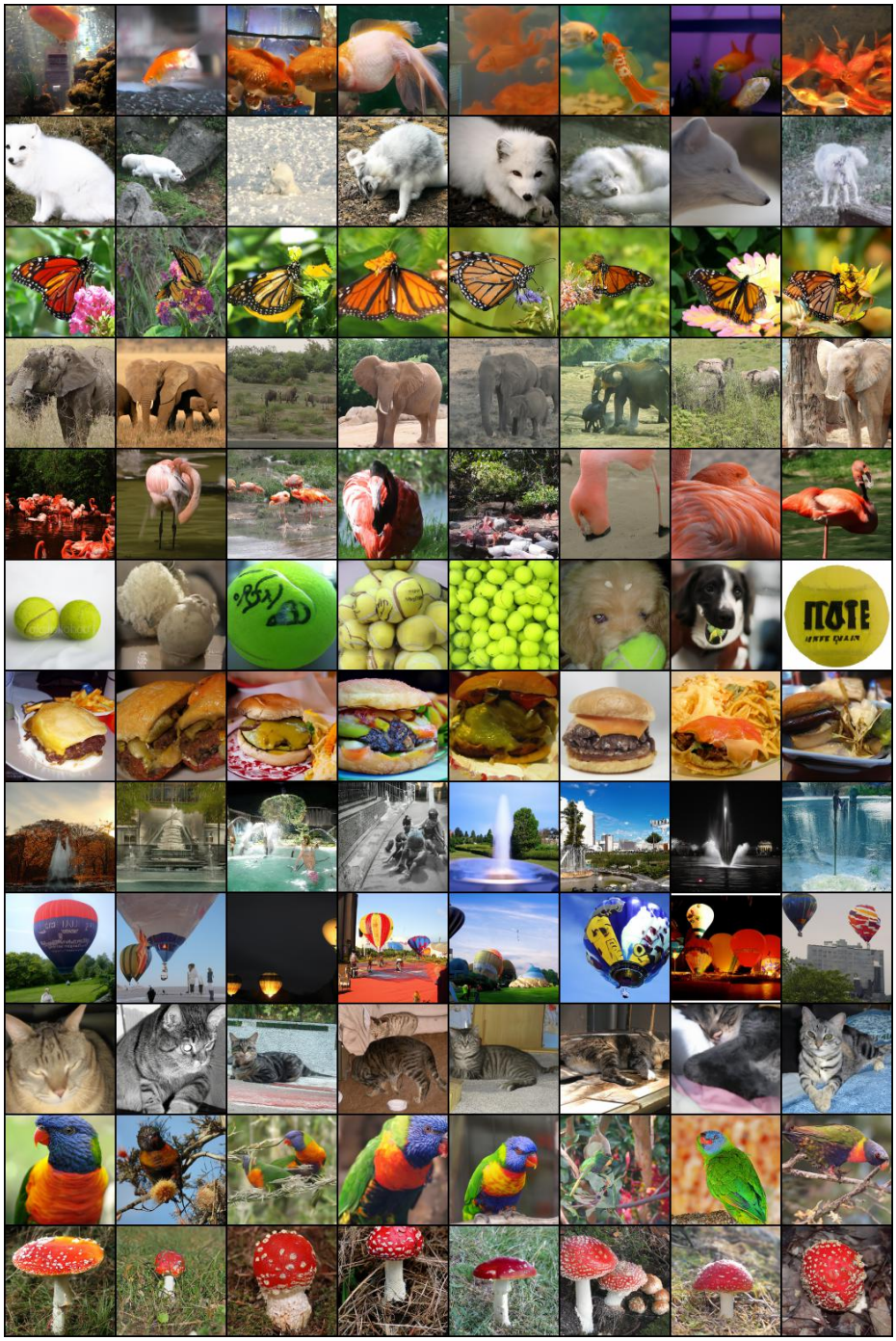


Fig. 5: Generated images in ImageNet 128×128 from CADM-G with EDS and ECT.

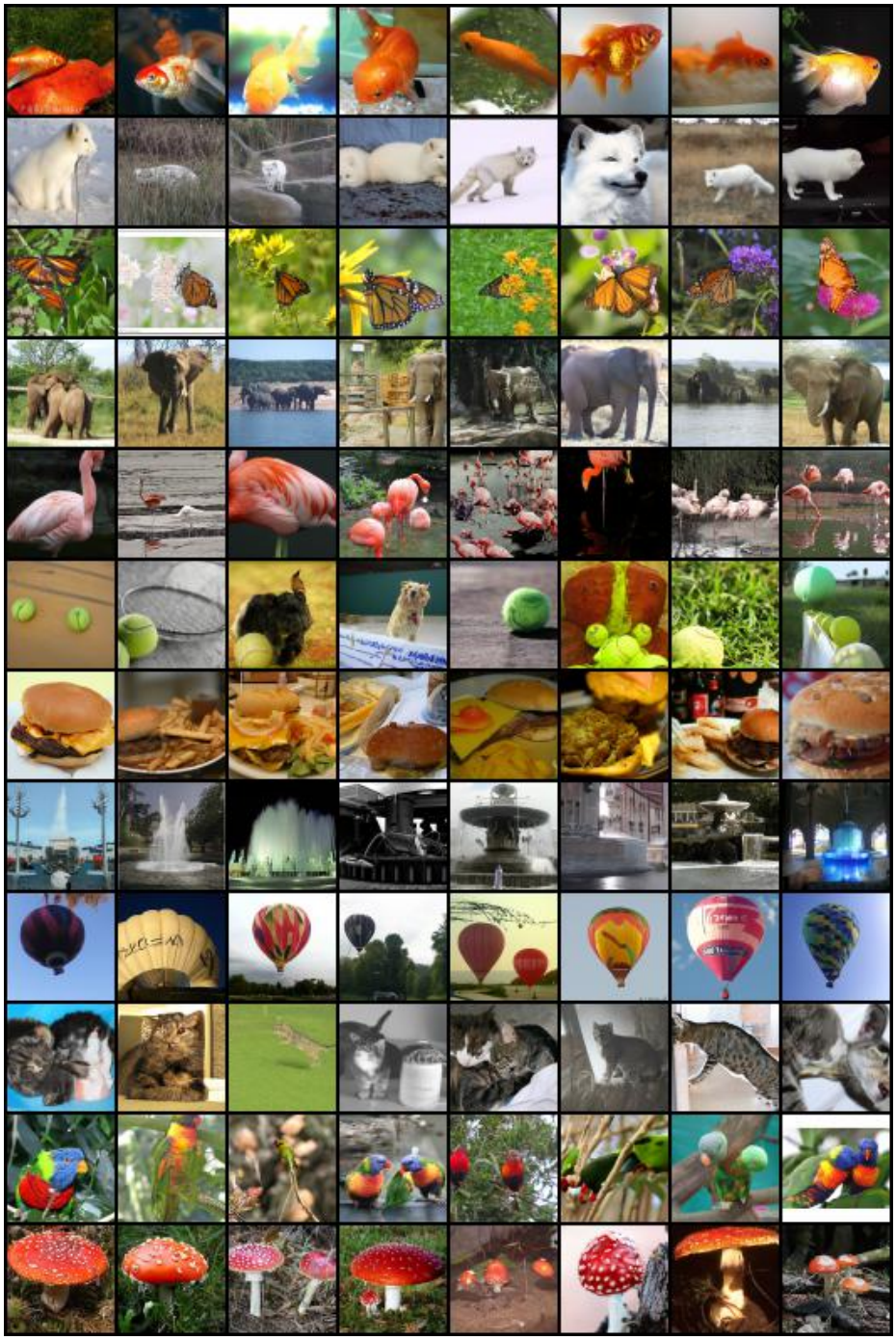


Fig. 6: Generated images in ImageNet 64×64 from CADM-G with EDS and ECT.

5 Ablation Study

5.1 Effectiveness of Hyperparameter γ.

In this section, we show the curve about the selection of hyperparameter γ. For efficient evaluation, we collect 5000 generated images and calculate FID metric based on 10000 real images from ImageNet1000.

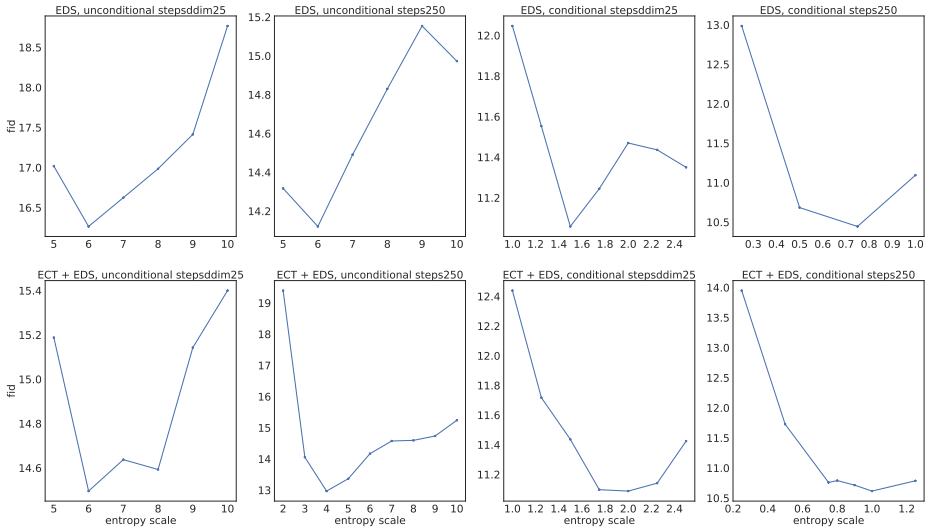


Fig. 7: The curve plot of optimal EDS hyperparameter γ based on 5k images evaluated with FID.

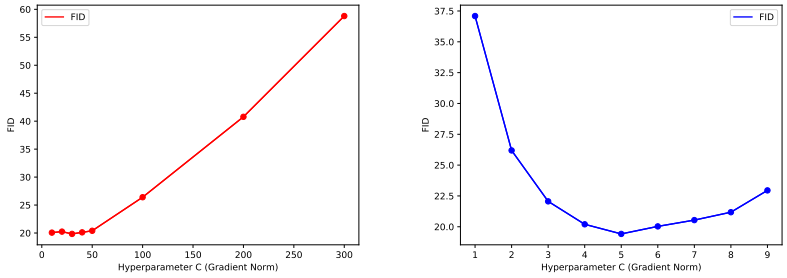
5.2 Effectiveness of intuitive sampling schemes.

An intuitive solution for sampling scheme is to manually select vanishing time point, i.e., 700 and finetune the constant rescaling factor to adjust the weak gradient for all generated samples, which we called Constant (range 0-700):

$$\mathbf{g} = \begin{cases} \nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y}|\mathbf{x}_t), & t > 700 \\ C * \nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y}|\mathbf{x}_t), & \text{Otherwise} \end{cases} \tag{1}$$

Another intuitive scaling design mentioned in main paper is to rescale the gradient guidance according to current time step t. When t is close to T, the scaling effect should be insignificant. Thus, time-aware gradient guidance can be formulated as following:

$$\mathbf{g} = C * (T - t) * \nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y}|\mathbf{x}_t), \tag{2}$$



(a) Ablation of Constant(range 0-700) (b) Ablation of Time-aware

Fig. 8: The curve plot of optimal EDS hyperparameter C based on 5000 images evaluated with FID for Constant and Time-aware methods separately.

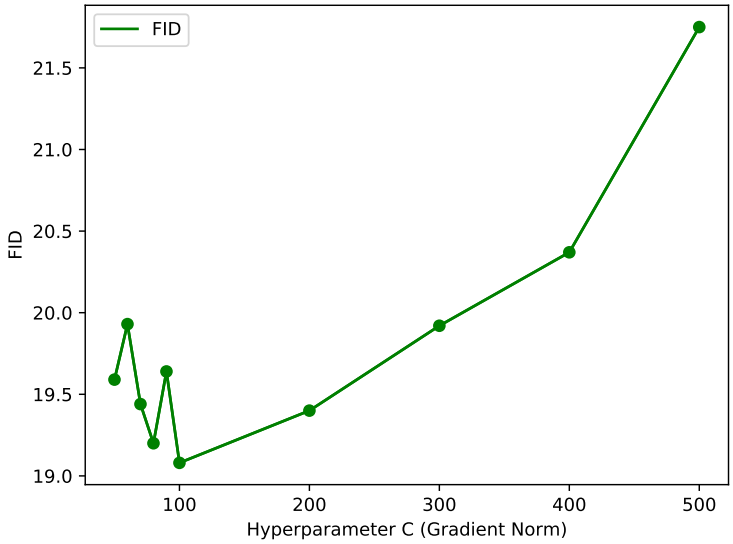


Fig. 9: Ablation of Gradient Norm

Fig. 10: The curve plot of optimal hyperparameter C based on 5k images for Gradient Norm method.

where C is the constant hyperparameter to balance the scaling and gradient effects. We call this method as Timestep-aware, which can dynamically adjust the scaling factor according to time step in sampling process. The ablations about the hyperparameter C for two methods above are shown in Fig. 8.

We design another approach which is based on norm of gradient map to adaptively adjust the gradient scale. Specifically, we empirically select a norm bound M, i.e., 0.2 for gradient. When the norm of gradient map is smaller than the threshold vanishing norm bound M, we regard that the gradient guidance is weak and need to be rescaled. Thus, the gradient term is rewritten as followed:

$$\mathbf{g} = \begin{cases} \nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y}|\mathbf{x}_t), & \|\nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y}|\mathbf{x}_t)\|_2 < 0.2 \\ C * \nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y}|\mathbf{x}_t), & \text{Otherwise} \end{cases} \tag{3}$$

where C is the constant hyperparameter to balance the scaling and gradient effects. The ablation about C in this method can be seen in Fig. 10.

References

1. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34** (2021)
2. Donahue, J., Simonyan, K.: Large scale adversarial representation learning. arXiv:1907.02544 (2019)
3. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
4. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models. arXiv preprint arXiv:2102.09672 (2021)
5. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models. arXiv:2102.09672 (2021)
6. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv:2011.13456 (2020)
7. Wu, Y., Donahue, J., Balduzzi, D., Simonyan, K., Lillicrap, T.: Logan: Latent optimisation for generative adversarial networks. arXiv:1912.00953 (2019)