Analyzing Stability in Diffusion Model Training: Empirical Analysis and Strategic Improvements

Ruilin Wang 31520231154317,[1] Zhijie Wei 36920231153239,[2] Zhanhao Xie 31520231154322,[1] Kaipeng Huang 31520231154290,[2] Yongcun Zhang 31520231154327,[1]

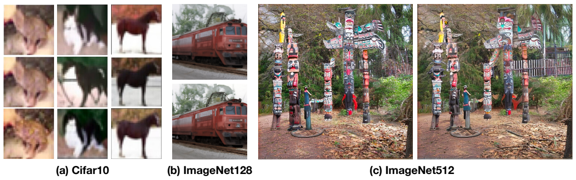
[1]Deep Learning AI Class
[2]Deep Learning Information School Class

Abstract

Diffusion models, a potent generative framework, have recently garnered substantial attention. While many argue that the advantages of diffusion models stem from their comparatively steady training procedure when contrasted with Generative Adversarial Networks (GANs), these assertions often rely on intuition and lack concrete empirical validation. This research paper aims to furnish direct empirical proof elucidating the impressive steadiness demonstrated by diffusion models during their training phase. Our methodology entails an inquiry that initiates a comparative examination of the generative results of models with differing hyperparameters, encompassing initialization and model architecture, under identical sampling circumstances. Our discoveries illustrate that diffusion models consistently produce uniform generative outcomes across various hyperparameter configurations, emphasizing their resilience in learning the association between random variations and data. Subsequently, we proceed to examine and compare the loss landscapes of diffusion models and GANs, disclosing that diffusion models exhibit notably smoother loss terrains, implying heightened convergence stability. Based on these experiments, we have conclusively validated the advantages of the diffusion model in terms of its architectural structure. Furthermore, employing the curriculum learning-based timestep schedule, we proposed a training optimization technique based on the principle of reducing difficulty from easy to hard, yielding nearly a twofold time optimization on CIFAR-10.

Introduction

Diffusion Models (DMs)(Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Song et al. 2020) , a prominent class of generative models, have received significant attention in recent years due to their exceptional ability to model complex data distributions. DMs have led to substantial advancements in various domains, including image generation(Nichol and Dhariwal 2021; Dhariwal and Nichol 2021; Rombach et al. 2021), image manipulation(Zhang, Rao, and Agrawala 2023; Lugmayr et al. 2022; Kawar et al. 2023), video generation(Ho et al. 2022; Blattmann et al. 2023; Wang et al. 2023), and speech synthesis(Jeong et al. 2021; Zhang et al. 2023). While the superior performance of diffusion models is often attributed to their stable training process, these claims are frequently based on intuition and lack empirical evidence.



(a) Cifar10 (b) ImageNet128 (c) ImageNet512

Figure 1: Illustration of the consistency phenomenon in diffusion models (DMs). Despite different initializations or structural variations, DMs trained on the same dataset produce remarkably consistent results when exposed to identical noise during sampling. (a) presents three models trained on CIFAR10 with different initializations. (b) depicts two models (Dhariwal and Nichol 2021) trained on ImageNet128 with different structures. (c) showcases the large and huge models of UViT (Bao et al. 2023) trained on ImageNet512.

In this study, we endeavor to provide empirical evidence substantiating the stability of the training process in DMs. Based on our findings, changing the model structure and initialization would not significantly influence the result as long as training with constant noise, *i.e.*, the same initial noise and noise per round. We try to reveal the veil of the stability of DMs from the perspective of the training landscape. Our results reveal a notable consistency in the generative outcomes, as depicted in Figure 1. It is important to highlight that such a consistency phenomenon is not typically observed in generative models. These models generally bootstrap samples that adhere to a certain probability distribution, *i.e.,* a noise, onto the desired data distribution in an ultra-high dimensional space (Song and Ermon 2019). This process is inherently laden with a considerable degree of randomness. Consequently, this experiment demonstrates the stability of DMs in learning noise-data mapping relationships and hyper-parameters robustness.

From the observation of DMs mentioned above, we can reasonably speculate that *The landscape of DMs resembles a bowl*, which implies that the model from different initialization and structure converges to a similar minimum. To further investigate the loss landscape associated with DMs, we

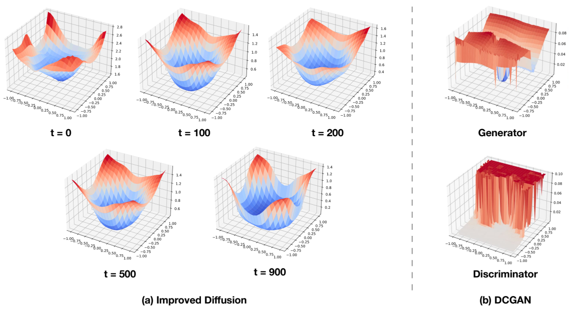


Figure 2: Visualization of the loss landscapes of Improved Diffusion and DCGAN, where **t** is the timestep of DMs. Both models were trained on the CIFAR10 dataset. The loss landscape of DMs is smoother compared to GANs.

employ various techniques such as loss landscape visualization (Li et al. 2018), 1D interpolation, and Hessian spectral decomposition (Yao et al. 2020). Loss landscape is the high-dimensional space formed by the partial derivatives of the loss function concerning the model parameters. The smoothness of this space affects the convergence rate of the model (Garrigos and Gower 2023), *i.e.,* the smoother the loss landscape, the easier the optimization. We compare the results of these techniques on DMs and a conventional generative model, *e.g.,* GANs, in Fig 2, respectively. Our findings reveal that the loss landscape of the diffusion model exhibits significantly higher smoothness compared to that of GANs, implying that DMs are easier to optimize than GANs.

Motivated by these analyses, we have proposed an optimization approach for DMs. Further investigation into the consistency phenomenon of DMs revealed varying convergence difficulties among different timesteps, each contributing differently to the final quality of generation (Choi et al. 2022). Hence, we introduced the curriculum learning-based timestep schedule (CLTS) (Bengio et al. 2009). This approach aims to gradually reduce the sampling probability of easily converging timesteps, thereby enhancing training efficiency.

To validate its effectiveness, we conducted training comparisons on CIFAR-10. The results indicate that this approach achieves nearly twice the time optimization on CIFAR-10. Moreover, it corroborates that the consistency observed aligns with meaningful phenomena.

Our contributions are summarized as follows:

- We provide empirical evidence by conducting a consistency experiment and comparing the loss landscapes of diffusion models and GANs, to elucidate why diffusion models exhibit significant stability during training.
- We propose an optimization method for diffusion models, namely the curriculum learning-based timestep schedule (CLTS). This method optimizes the sampling probability of timesteps to expedite model convergence.
- We evaluate the effectiveness of our optimization methods on diffusion models and datasets, showcasing their impact on improving the convergence speed of diffusion models.

Related Work

Diffusion Models

Diffusion Models (DMs) are a class of generative models that use techniques from non-equilibrium thermodynamics to learn the latent structure of complex data distributions. They were first introduced by (Sohl-Dickstein et al. 2015), who applied their method to image and text generation. Later, (Ho, Jain, and Abbeel 2020) proposed Denoising Diffusion Probabilistic Models (DDPM), which improved the sampling efficiency and quality of diffusion models by using a denoising score matching objective and a learned diffusion process. (Rombach et al. 2021) developed Latent Diffusion Models (LDMs), which compressed high-resolution images into lower-dimensional representations using pre-trained autoencoders. They also introduced cross-attention layers into the model architecture, which enabled LDMs to handle various conditioning inputs, such as text or bounding boxes, and generate high-resolution images in a convolutional manner. Despite the success of DMs using UNet, a convolutional neural network, (Bao et al. 2023) and (Peebles and Xie 2022) discovered the feasibility of using Vision Transformer (Dosovitskiy et al. 2020) in DMs, achieving state-of-the-art generation results.

Several studies have focused on improving the DMs from various aspects (Karras et al. 2022; Chen 2023). (Nichol and Dhariwal 2021) proposed several techniques to enhance the performance and efficiency of DMs, such as employing a learned variance schedule, adopting a cosine timestep schedule for low-resolution data, and developing a multi-scale architecture. (Dhariwal and Nichol 2021) further improved the performance and fidelity of DMs, by incorporating advanced design concepts of BigGAN (Brock, Donahue, and Simonyan 2018). Although integrating the sophisticated model structure of GAN can benefit the performance of DMs, they also adopt the same large momentum setting, which is sub-optimal, because the loss landscape of DMs is highly smoothed. A large momentum not only affects convergence efficiency but also causes oscillations. This issue and its implications are explored in further detail in the following section.

Unveiling the Stability of Diffusion Models

In this section, we present empirical evidence to substantiate the stability of DMs in learning noise-to-data mapping and convergence, thereby underscoring their superiority over GANs. We delve into an analysis of the stability of DMs in the context of learning noise-to-data mapping. Finally, we draw a comparison between the smoothness of the loss landscape of DMs and GANs.

Analyze the Stability of noise-to-data mapping

The stability of the generative model learning the noise-to-data mapping is an important aspect of generative modeling, as it reflects how well the model can cope with different noise and different choices of hyper-parameters. However, this stability is often overlooked or not explicitly evaluated.

In this section, we evaluate the stability of DMs in learning the noise-to-data mapping through a consistency exper-

Table 1: Comparing consistencies of DMs and GANs in learning noise-to-data mapping.

	Datasets	Different Initializations	Different Structures	Consistency (PSNR)
Improved Diffusion	Cifar10	✓		**20.14**
DCGAN	Cifar10	✓		10.48
Guided Diffusion	ImageNet128	✓	✓	**17.23**
BigGAN	ImageNet128	✓	✓	8.58
U-ViT	ImageNet512	✓	✓	**14.37**
BigGAN	ImageNet512	✓	✓	6.40

iment. We select three diffusion frameworks as representative DMs: Improved Diffusion , Guided Diffusion, and U-ViT, and two GAN frameworks: DCGAN (Radford, Metz, and Chintala 2015) and BigGAN. We train DMs and GANs with different hyper-parameters on three benchmarks: Cifar10 (Krizhevsky, Hinton et al. 2009), ImageNet128 (Deng et al. 2009) and ImageNet512 (Deng et al. 2009). The results of the consistency experiment and the detailed settings are presented in Table 1.

In the consistency experiment, we use the peak signal-to-noise ratio (PSNR) to measure the consistency of each model. Specifically, for a group of models, *e.g.,* For different initializations of Improved Diffusion or large and huge models of U-ViT, we sample 32 images with the same sampling seed to ensure identical initial and noise in each round. Suppose we have N models of a group, each model generate M images, we then measure the consistency $C(\cdot)$,

$$C(q) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N-1}\sum_{j=2}^{N}\mathrm{PSNR}(q_{i,1},\ q_{i,j}), \qquad (1)$$

where $q \in \mathbb{R}^{N\times M}$ is the matrix of images.

The result of the consistency experiment reveals that all DMs have much higher consistency than GANs, regardless of the dataset, initialization, or model structure, indicating that DMs are more robust and stable in learning noise-to-data mapping.

Analyze the Smoothness of Landscape

The smoothness of the loss landscape is strongly correlated with the convergence difficulty. In this section, we conduct a thorough investigation of the loss landscape of DMs and GANs during the training. However, due to the high dimensionality of the models' parameters, it is infeasible to access the full information of the loss landscape. Therefore, we resort to a partial analysis based on 1D interpolation of models and hessian spectra, following the method proposed by (Li et al. 2018).

1D interpolation is a technique that generates new data points by leveraging existing data. In our research, we employed 1D linear interpolation to estimate the position $\boldsymbol{\theta}$ within the landscape using the provided models at different stages of training, namely $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$. This involved calculating the weighted sum of these two models.

$$\boldsymbol{\theta} = \alpha\boldsymbol{\theta}_a + (1-\alpha)\boldsymbol{\theta}_b. \quad (0 \le \alpha \le 1) \qquad (2)$$

We use interpolation to analyze the relationship between different training stages and gather valuable information. Our approach involves training a Diffusion model and a GAN model, followed by selecting models from various training steps as anchor points. Specifically, we select the models trained 10 and 100 epochs for both DM and GAN. These selections, shown in Figure 3, represent models from both early and late convergence stages.

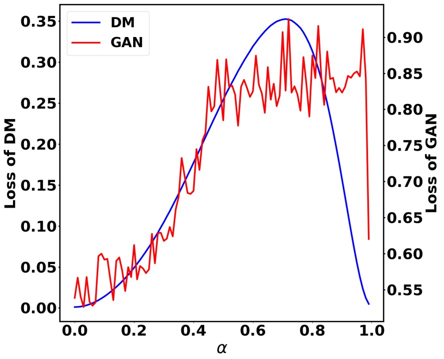


Figure 3: Illustration of the 1D-interpolation results of DMs and GANs. The jitter red line indicates the geometry of GAN's landscape is rougher.

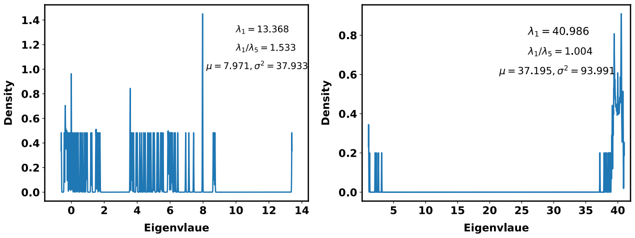


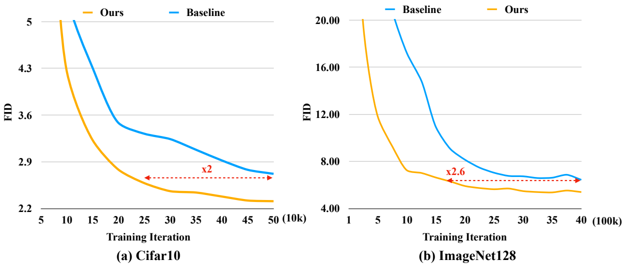
Figure 4: Illustration of the Hessian spectrum of DMs(left) and GANs(right). λ_i is the i-th largest eigenvalue and μ and σ is the mean and variance of eigenvalue respectively. Larger dominant eigenvalue, sharper the landscape, and the greater the differences among eigenvalues, the more difficult the model is to optimize.

As shown in Figure 3, the GAN model exhibits more erratic changes in loss, indicating that changes in GAN parameters lead to relatively larger changes.

Hessian spectra refers to the distribution of eigenvalues in the Hessian matrix. Inspired by the connection between the geometry of the loss landscape and the eigenvalue, we approximate the Hessian spectrum by the Lanczos algorithm and the results of Diffusion and GAN are shown in Figure 4. From the figure, it can be seen that the dominant eigenvalue of GAN's is larger, *i.e.*, $\lambda_1 = 13.3$(DMs) *v.s.* $\lambda_1 = 40.9$(GANs), and dispersion, *i.e.*, $\sigma^2 = 37.9$(DMs) *v.s.* $\sigma^2 = 93.9$(GANs), which implies that the landscape of GAN is steeper and more rugged, which also means that the GAN is more difficult to optimize.

Optimization

In this section, we delve into the underlying reasons for the unique consistency phenomenon observed exclusively



(a) Cifar10

(b) ImageNet128

Figure 5: Illustration of the application of our optimization approach on different DMs. (a) Improved Diffusion trained on Cifar10, (b) Guided Diffusion trained on ImageNet128. With our methods, these DMs achieve 2× and 2.6× speedup in training, respectively.

in DMs. We demonstrate that ϵ-predicted DMs tend to become trivial as timesteps approach T, resulting in high structural similarity but low diversity in image details. Subsequently, we explore leveraging this characteristic to optimize DMs, drawing inspiration from curriculum learning. We introduce a timestep schedule that gradually reduces the sampling probabilities of timesteps $t \rightarrow T$ as training progresses.

Investigating the Consistency Phenomenon

We initiate our exploration by formulating the forward diffusion process as follows:

$$x_t = \sqrt{\bar{\alpha}_t} \cdot x_0 + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \Pi_{s=0}^{t} \alpha_s$. Note that $\bar{\alpha}_t$ is a factor ranging from 0 to 1. As $t \rightarrow T$, $\bar{\alpha}_t \rightarrow 0$, leading to $x_t \rightarrow \epsilon$.

The simplicity loss, denoted as L_{simple}, is defined as:

$$L_{\text{simple}} = E_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,\mathbf{I})} \left[||\epsilon - \epsilon_\theta(x_t, t)||^2 \right], \tag{4}$$

which, as $x_t \rightarrow \epsilon$, results in $\epsilon_\theta \rightarrow \mathbf{I}$. This indicates that ϵ-predicted DMs tend to become trivial as $t \rightarrow T$.

Results from consistency experiments (Fig. 1) support the aforementioned derivation. Images exhibiting the consistency phenomenon are structurally similar yet differ in detail, especially when the diffusion model generates structural information at $t \rightarrow T$.

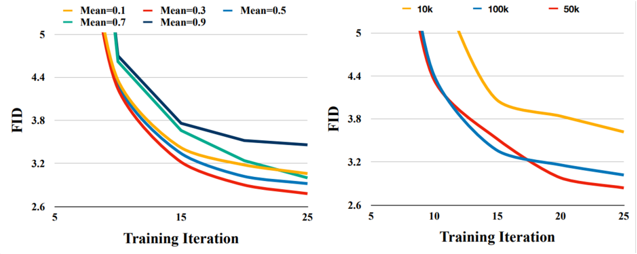
To further confirm that the ϵ-predicted mechanism leads to model triviality as $t \rightarrow T$, we trained an x_0-predicted DM as a counterexample. We modified the loss function to:

$$L_{x_0} = E_{x_0 \sim q(x_0)} \left[||x_0 - \mu_\theta(x_t, t)||^2 \right]. \tag{5}$$

We conclude that the ϵ-predicted mechanism is the root cause of the consistency phenomenon as $t \rightarrow T$.

Optimizing Sampling Probabilities of Timesteps in Training

The presence of the consistency phenomenon suggests that DMs tend to converge easily as timesteps approach T. Consequently, we propose an innovative approach to enhance the training efficiency of DMs. Notably, we observe that existing diffusion frameworks treat all timesteps equally, employing a uniform probability $U(t) = 1/T$ for timestep sampling during training. This results in redundant training for $t \rightarrow T$.



(a)

(b)

Figure 6: Ablation study, every model is trained on Cifar10. (a) illustrates the influence of different mean μ in our proposed CLTS (Eq. 6). (b) reflects the influence of values of different target iterations that we used in CLTS (Eq. 7).

To tackle this challenge, we embrace curriculum learning, a training acceleration technique grounded in the principle of learning from easy to hard. Interestingly, DMs inherently generate data with varying levels of difficulty, with the difficulty of ϵ-predicted DMs increasing as the timestep decreases.

Our solution is the Curriculum Learning-based Timestep Schedule (CLTS), designed to progressively decrease the sampling probabilities of timesteps $t \rightarrow T$ as training progresses while increasing the probabilities of others—essentially finding an optimal timestep distribution. For simplicity, we assume that the optimal timestep distribution follows a Gaussian distribution $N(\cdot)$, where the mean μ signifies the most important timesteps, and others are less critical:

$$N(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right). \tag{6}$$

To streamline CLTS and reduce hyper-parameter complexity, we set the variance $\sigma = T$, establishing a standard variance across timesteps. Our initial attempt involved shifting the Gaussian distribution as the mean transitions from T to 0, but this yielded minimal improvement. We hypothesize that the generation of DMs requires the involvement of all timesteps. Shifting the distribution led to overly small sampling probabilities for all timesteps except $t \rightarrow T$ at the initial stage. Therefore, we propose a mixed distribution, introducing a factor γ to transition the distribution from uniform to Gaussian:

$$P(t) = (1-\gamma)U(t) + \gamma N(t), \; \gamma = \frac{\text{current iteration}}{\text{target iteration}}, \tag{7}$$

where the target iteration is a hyper-parameter that controls the speed of the Gaussian distribution's emergence.

It is important to note that our proposed CLTS shares a similar implementation with (Hang et al. 2023) and (Choi et al. 2022). However, there are substantial differences in our underlying philosophies. Inspired by curriculum learning, our approach is rooted in the principle of learning from easy to hard, while (Hang et al. 2023) and (Choi et al. 2022) focus on finding an optimal distribution. Our method demonstrates increased robustness and efficiency in extensive experiments, as showcased in the next section.

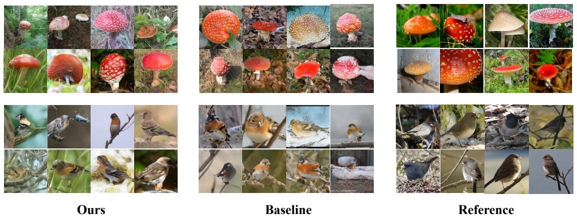


Figure 7: Comparisons of generated images. Both ours and the baseline are trained on Guided Diffusion

Experiments

In this section, we train our optimized methods on cifar10 and ImageNet128 datasets, following the hyper-parameter settings of Improved Diffusion and Guided Diffusion, respectively. The details of the hyper-parameter settings are as follows:

For Cifar10, we used a cosine timestep schedule, 4,000 timesteps, learning rate = 1e-4, and batch size = 128. We used an exponential moving average (EMA) rate of 0.9999 for all experiments. We implemented our models in PyTorch, and trained them on 2 × NVIDIA 3090 GPUs, using 250 sampling processes. We used Adam optimizer, with β_1 = 0.8, β_2 = 0.999. which are based on the observation of the smooth landscape of diffusion models (DMs). For CLTS, we set the mean value μ = 1200 (0.3 × total timesteps, the optimized mean value through ablation study), and the target iteration = 5×10^4.

For ImageNet128, we used a linear timestep schedule, 1,000 timesteps, learning rate = 1e-4, and batch size = 256. We also used an EMA rate of 0.9999 for all experiments. We implemented our models in PyTorch, and trained them on 2 × NVIDIA A800 GPUs, using 250 sampling processes. We used Adam optimizer, with initial β_1 = 0.8, β_2 = 0.999. The hyper-parameters of our proposed methods are as follows: For CLTS, we set the mean value μ = 300, and and the target iteration = 3×10^5.

Ablations

To evaluate the effectiveness of our proposed methods, we performed an ablation study. The results indicate that each module enhances the performance of the model, and the combination of all modules achieves the best FID score. Fig. 6 (a) examines the effect of different mean values μ in our proposed CLTS (Eq. 6). The mean value μ controls the most important timesteps in the Gaussian distribution. The results suggest that the optimal value of μ is around 0.3, which implies that the timesteps with the highest contribution to generation are not necessarily the most difficult ones to learn. Fig. 6 (b) investigates the effect of different target iterations in our proposed CLTS (Eq. 7). The target iteration is a hyper-parameter that adjusts the speed of the Gaussian distribution emerging. The results demonstrate that the optimal value of the target iteration is around 100k, which means that the model needs about 100k iterations to fully adapt to the Gaussian distribution.

Table 2: Comparing with state-of-the-art methods in ImageNet128, we use FID to evaluate the performance. Our method achieves the lowest FID score at each iteration.

Methods	Iters=1M	Iters=2M	Iters=3M	Iters=4M
GD	17.18	8.14	6.63	6.04
Min-SNR	13.53	6.49	6.11	5.81
GD+Ours	**7.24**	**5.91**	**5.48**	**5.40**

Table 3: Comparing with state-of-the-art methods in Cifar10, we use FID to evaluate the performance. Our method achieves the lowest FID score at each iteration.

Methods	Iters=100k	Iters=200k	Iters=300k	Iters=400k	Iters=500k
ID	5.40	3.48	3.05	2.72	2.60
FDM	4.91	3.03	2.58	2.49	2.43
ID+Ours	**4.24**	**2.81**	**2.46**	**2.38**	**2.31**

Based on the optimal settings, we trained our optimized models on Cifar10 and ImageNet 128, and compared them with the baseline models. Fig. 5 illustrates the results. The results reveal a significant acceleration of our optimized models, *e.g.,* on Cifar10, our model achieves a 2× speedup compared with the baseline model, and on ImageNet128, our model achieves a 2.6× acceleration. Fig. 7 shows the visualization results of our methods. These results demonstrate the effectiveness and robustness of our proposed methods.

Comparisons with state-of-the-art methods

We compare our optimized models with state-of-the-art methods, Min-SNR and FDM (Wu et al. 2023). Table 2 compares the performance of our method with two state-of-the-art methods, Guided Diffusion (GD) and Min-SNR, on ImageNet128, and Table 3 compares with Improved Diffusion (ID) and FDM (Wu et al. 2023), on Cifar10. The results demonstrate that our method achieves the lowest FID score at each iteration of both datasets, indicating that our method outperforms the existing methods in terms of image generation quality and speed.

Conclusion

In this paper, we have investigated the consistency phenomenon of diffusion models (DMs). We have attributed this phenomenon to two factors: the lower learning difficulty of DMs at higher noise rates, and the smoothness of the loss landscape of DMs. Based on this finding, we have proposed the strategy to accelerate the training of DMs: a curriculum learning based timestep schedule. We have evaluated our proposed strategies on various models and datasets, and demonstrated that they can significantly reduce the training time and improve the quality of the generated images. Our work not only reveals the stability of DMs, but also provides practical guidance for training DMs more efficiently and effectively.

References

Bao, F.; Nie, S.; Xue, K.; Cao, Y.; Li, C.; Su, H.; and Zhu, J. 2023. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22669–22679.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *International Conference on Machine Learning*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *ArXiv*, abs/1809.11096.

Chen, T. 2023. On the Importance of Noise Scheduling for Diffusion Models. *ArXiv*, abs/2301.10972.

Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H. J.; and Yoon, S.-H. 2022. Perception Prioritized Training of Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11462–11471.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

Garrigos, G.; and Gower, R. M. 2023. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*.

Hang, T.; Gu, S.; Li, C.; Bao, J.; Chen, D.; Hu, H.; Geng, X.; and Guo, B. 2023. Efficient Diffusion Training via Min-SNR Weighting Strategy. *ArXiv*, abs/2303.09556.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jeong, M.; Kim, H.; Cheon, S. J.; Choi, B. J.; and Kim, N. S. 2021. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409*.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. *ArXiv*, abs/2206.00364.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Peebles, W. S.; and Xie, S. 2022. Scalable Diffusion Models with Transformers. *ArXiv*, abs/2212.09748.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023. VideoComposer: Compositional Video Synthesis with Motion Controllability. *arXiv preprint arXiv:2306.02018*.

Wu, Z.; Zhou, P.; Kawaguchi, K.; and Zhang, H. 2023. Fast Diffusion Model. *ArXiv*, abs/2306.06991.

Yao, Z.; Gholami, A.; Keutzer, K.; and Mahoney, M. W. 2020. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, 581–590. IEEE.

Zhang, C.; Zhang, C.; Zheng, S.; Zhang, M.; Qamar, M.; Bae, S.-H.; and Kweon, I. S. 2023. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai. *arXiv preprint arXiv:2303.13336*, 2.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.