

# Non-uniform Timestep Sampling: Towards Faster Diffusion Model Training Supplementary Materials

Tianyi Zheng\*  
Shanghai Jiao Tong University  
Shanghai, China  
tyzheng@sjtu.edu.cn

Cong Geng  
vivo Mobile Communication Co., Ltd  
Shanghai, China  
gengcong@vivo.com

Peng-Tao Jiang  
vivo Mobile Communication Co., Ltd  
Shanghai, China  
pt.jiang@vivo.com

Ben Wan  
Shanghai Jiao Tong University  
Shanghai, China  
burn-w@sjtu.edu.cn

Hao Zhang  
vivo Mobile Communication Co., Ltd  
Shanghai, China  
haozhang@vivo.com

Jinwei Chen  
vivo Mobile Communication Co., Ltd  
Shanghai, China  
jinwei.chen@vivo.com

Jia Wang†  
Shanghai Jiao Tong University  
Shanghai, China  
jiawang@sjtu.edu.cn

Bo Li†  
vivo Mobile Communication Co., Ltd  
Shanghai, China  
libra@vivo.com

## CCS Concepts

• Computing methodologies → Maximum likelihood modeling; Reconstruction.

## Keywords

Diffusion Model, Optimal Transport, Wasserstein distance.

## ACM Reference Format:

Tianyi Zheng, Cong Geng, Peng-Tao Jiang, Ben Wan, Hao Zhang, Jinwei Chen, Jia Wang, and Bo Li. 2024. Non-uniform Timestep Sampling: Towards Faster Diffusion Model Training Supplementary Materials. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28–November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3664647.3680912>

## 1 Overview

In the appendix, we first provide detailed proof of the previous Lemma, Theorem, and Corollary in Section 2. Then we provide more experiment results and analysis in Section 3. In Section 4, we show the details and network architecture. Finally, we present more visual results in Section 5.

## 2 The detailed proof.

### 2.1 Proof for Lemma 1

**Lemma 1.** Consider a regular Wasserstein gradient flow, as defined in Equation 3 in the manuscript, initiating from a data distribution

\*This work was done during Tianyi Zheng’s internship at vivo.

†Corresponding Authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0686-8/24/10

<https://doi.org/10.1145/3664647.3680912>

$\mu_0$  and converging to a normal Gaussian distribution  $\mu_T = \mathcal{N}(0, I)$ . With the selection of  $f = \beta_t x$  and  $g_t = \sqrt{2\beta_t}$ , the family of measures  $\{\mu_t\}_{t=0}^T$  derived from the Fokker-Planck equation Equation 4 in the manuscript is equivalent to the family of measures corresponding to this gradient flow.

**Proof of Lemma 1.** The Fokker-Planck Equation in the manuscript is

$$\frac{\partial \mu_t}{\partial t} = -\nabla \cdot (\mu_t f) + \frac{1}{2} \nabla \cdot \left( g_t^2 \mu_t \right). \quad (1)$$

This Equation is equivalent to the following:

$$\begin{aligned} \frac{\partial \mu_t}{\partial t} &= -\nabla \cdot (\mu_t f) + \nabla \cdot \left( \frac{1}{2} \mu_t g_t^2 \nabla \log(\mu_t) \right) \\ &= -\nabla \cdot \left( \mu_t \left( f - \frac{1}{2} g_t^2 \nabla \log(\mu_t) \right) \right). \end{aligned} \quad (2)$$

We recall that the regular Wasserstein gradient flow [4] in the manuscript is

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot \left[ \mu_t \beta_t \nabla_{W_2} \mathcal{F} \right]. \quad (3)$$

Since the functional on the Wasserstein space is defined by  $\mathcal{F}(\mu) = \text{KL}(\cdot \| \pi)$ , therefore, we have:

$$\nabla_{W_2} \text{KL}(\cdot \| \pi) = \nabla \log \left( \frac{\mu}{\pi} \right).$$

Then, the Equation 3 can written as:

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot \left[ \mu_t (\beta_t \nabla \log(\pi) - \beta_t \nabla \log(\mu_t)) \right]. \quad (4)$$

Since  $\pi = \mu_T = \mathcal{N}(0, I) = \exp(-\frac{\|x\|^2}{2})/Z$ , When we set  $f = -\beta_t x$  and  $g_t = \sqrt{2\beta_t}$ . The Equation 2 is the same as Equation 4. Therefore, the family of measures  $\{\mu_t\}_{t=0}^T$  derived from the Fokker-Planck equation [18] Equation 4 in the manuscript is equivalent to the family of measures corresponding to this gradient flow. Meanwhile, the corresponding stochastic differential equation (SDE) of the Fokker-Planck Equation is:

$$dx = -\beta_t x dt + \sqrt{2\beta_t} dw. \quad (5)$$

This SDE is exactly the forward diffusion process of the diffusion model [6, 20].

## 2.2 Proof for Theorem 1

We first introduce the Grönwall's lemma [5].

**Lemma 2.** Let  $I$  denote an interval of the real line of the form  $[a, \infty)$  or  $[a, b]$  or  $[a, b)$  with  $a < b$ . Let  $\beta$  and  $u$  be real-valued continuous functions defined on  $I$ . If  $u$  is differentiable in the interior  $I^\circ$  of  $I$  (the interval  $I$  without the end points  $a$  and possibly  $b$ ) and satisfies the differential inequality

$$u'(t) \leq \beta(t)u(t), \quad t \in I^\circ \quad (6)$$

then  $u$  is bounded by the solution of the corresponding differential equation  $v'(t) = \beta(t)v(t)$

$$u(t) \leq u(a) \exp\left(\int_a^t \beta(s)ds\right). \quad (7)$$

for all  $t \in I$ .

**Proof of Lemma 2.** Define the function

$$v(t) = \exp\left(\int_a^t \beta(s)ds\right), \quad t \in I. \quad (8)$$

Note that  $v$  satisfies

$$v'(t) = \beta(t)v(t), \quad t \in I^\circ. \quad (9)$$

with  $v(a) = 1$  and  $v(t) > 0$  for all  $t \in I^\circ$ . By the quotient rule

$$\begin{aligned} \frac{d}{dt} \frac{u(t)}{v(t)} &= \frac{u'(t)v(t) - v'(t)u(t)}{v^2(t)} \\ &= \frac{u'(t)v(t) - \beta(t)v(t)u(t)}{v^2(t)} \leq 0, \quad t \in I^\circ. \end{aligned} \quad (10)$$

Thus the derivative of the function  $u(t)/v(t)$  is non-positive and the function is bounded above by its value at the initial point  $a$  of the interval  $I$ .

$$\frac{u(t)}{v(t)} \leq \frac{u(a)}{v(a)} = u(a), \quad t \in I. \quad (11)$$

which is Grönwall's lemma.

**Theorem 1.** Consider two distinct initial distributions  $\mu_0$  and  $\hat{\mu}_0$  on the data manifold  $M$ , which is equipped with a reference measure  $\nu = e^{-V} \text{vol}$ , satisfying  $\text{Hess}_\mu \geq K$ . Let  $\mu_t$  and  $\hat{\mu}_t$  represent the distributions at time  $t$  in the forward diffusion process, originating from  $\mu_0$  and  $\hat{\mu}_0$  respectively. For all  $t > 0$ , the following inequality holds

$$W_2(\mu_t, \hat{\mu}_t) \leq e^{-Kt} W_2(\mu_0, \hat{\mu}_0), \quad (12)$$

**Proof of Theorem 1.** Based on previous analyze,  $\mu_t$  and  $\hat{\mu}_t$  are two gradient flow. For fixed  $t$ , assume  $\mu_t^s$  is a geodesic in  $\mathcal{P}_2(M)$  joining  $\mu_t^0 = \mu_t$  and  $\mu_t^1 = \hat{\mu}_t$ . Using Taylor's theorem with the integral form of remainder.

$$U_\nu(\mu_t^1) - U_\nu(\mu_t^0) = U'_\nu(\mu_t^0) + \int_0^1 (1-s)U''_\nu(\mu_t^s) ds. \quad (13)$$

Then according to the identity that  $\text{Hess}_{\mu_t} U_\nu(\mu_t) = \frac{d^2 U_\nu(\mu_t^s)}{ds^2}$  (be cautious that  $\mu_t$  is with respect to  $s$ ), and the assumption of  $\text{Hess}_{\mu_t} \geq K$ ,

$$\begin{aligned} U_\nu(\mu_t^1) - U_\nu(\mu_t^0) &\geq U'_\nu(\mu_t^0) + K \int_0^1 (1-s) \|\dot{\mu}_t^s\|_{\mu_t^s}^2 ds \\ &= \langle \text{grad } U_\nu(\mu_t^s), \dot{\mu}_t^s \rangle_{s=0} + K \int_0^1 (1-s) \|\dot{\mu}_t^s\|_{\mu_t^s}^2 ds \\ &= \langle \text{grad } U_\nu(\mu_t^s), \dot{\mu}_t^s \rangle_{s=0} + \frac{K}{2} W_2^2(\mu_t^0, \mu_t^1), \end{aligned} \quad (14)$$

since  $\mu_t^s$  is a geodesic. Similarly, we have:

$$U_\nu(\mu_t^0) - U_\nu(\mu_t^1) \geq -\langle \text{grad } U_\nu(\mu_t^s), \dot{\mu}_t^s \rangle_{s=1} + \frac{K}{2} W_2^2(\mu_t^0, \mu_t^1). \quad (15)$$

Then by adding these inequalities,

$$\langle \text{grad } U_\nu(\mu_t^s), \dot{\mu}_t^s \rangle_0^1 \geq KW_2^2(\mu_t^0, \mu_t^1) \quad (16)$$

Now we calculate:

$$\begin{aligned} \frac{d}{dt} W_2^2(\mu_t, \hat{\mu}_t) &= \frac{d}{dt} \int_0^1 \|\dot{\mu}_t^s\|^2 ds \\ &= \int_0^1 \frac{d}{dt} \left\langle \frac{\partial}{\partial s} \mu_t^s, \frac{\partial}{\partial s} \mu_t^s \right\rangle ds \\ &= 2 \int_0^1 \left\langle \frac{D}{dt} \frac{\partial}{\partial s} \mu_t^s, \frac{\partial}{\partial s} \mu_t^s \right\rangle ds \\ &= 2 \int_0^1 \left\langle \frac{D}{ds} \frac{\partial}{\partial t} \mu_t^s, \frac{\partial}{\partial s} \mu_t^s \right\rangle ds \\ &= 2 \int_0^1 \left[ \frac{d}{ds} \left\langle \frac{\partial}{\partial t} \mu_t^s, \frac{\partial}{\partial s} \mu_t^s \right\rangle - \left\langle \frac{\partial}{\partial t} \mu_t^s, \frac{D}{ds} \frac{\partial}{\partial s} \mu_t^s \right\rangle \right] ds \\ &= 2 \int_0^1 \frac{d}{ds} \left\langle \frac{\partial}{\partial t} \mu_t^s, \frac{\partial}{\partial s} \mu_t^s \right\rangle ds \\ &= 2 \left\langle \frac{\partial}{\partial t} \mu_t^s, \frac{\partial}{\partial s} \mu_t^s \right\rangle_0^1 \\ &= -2 \langle \text{grad } U_\nu(\mu_t^s), \dot{\mu}_t^s \rangle_0^1 \\ &\leq -2KW_2^2(\mu_t, \hat{\mu}_t). \end{aligned} \quad (17)$$

Then we finish the proof by using Grönwall's lemma.

## 2.3 Proof for Proposition 3

**Proposition 3.** During the training stage, let  $X_n$  be a random variable representing a sample at time  $t_n$  in the timestep series  $\{t_1, t_2, \dots, t_\delta, \dots, t_T\}$ . Specifically,  $\Pr[X_n = 1]$  denotes the probability of sampling within the significant intervals at time  $n$ . The probability of having exactly  $k$  samples in the significant intervals, denoted by  $\Pr[S_n = k]$ , is given by the binomial formula  $\binom{n}{k} p^k (1-p)^{n-k}$ , consistent with a Bernoulli distribution.

**Proof of Proposition 3.** In the training stage of the diffusion model, each time is sampled from an independent and identically distributed (i.i.d.) The moment generating function of  $S_n$  is:

$$\begin{aligned} E \left[ e^{tS_n} \right] &= E \left[ e^{t(X_1 + \dots + X_n)} \right] = \prod_{j=1}^n E \left[ e^{tX_j} \right] \\ &= (pe^t + (1-p))^n = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{(n-k)} e^{tk} \end{aligned} \quad (18)$$

Since all moments of  $S_n$  exist ( $E \left[ |S_n|^k \right] \leq n^k < \infty$ ) and

$$\begin{aligned} \left| \sum_{k=0}^{\infty} \frac{t^k}{k!} E \left[ S_n^k \right] \right| &\leq \sum_{k=0}^{\infty} \frac{|t|^k}{k!} E \left[ |S_n|^k \right] \\ &\leq \sum_{k=0}^{\infty} \frac{|t|^k}{k!} n^k \\ &= \sum_{k=0}^{\infty} \frac{|tn|^k}{k!} \\ &= e^{|tn|} < \infty \end{aligned} \quad (19)$$

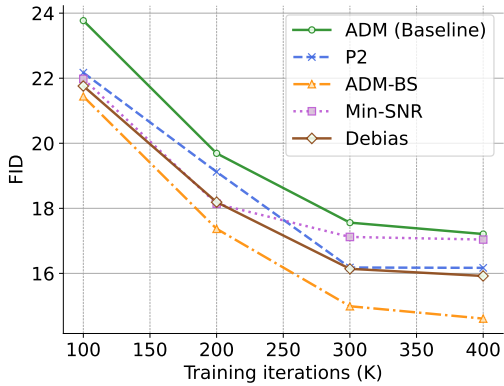
Therefore, by Theorem 30.1 in [1], the distribution of  $S_n$  shall equal

$$\Pr[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}. \quad (20)$$

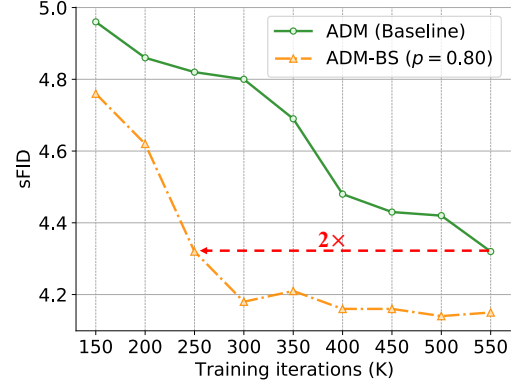
### 3 Additional results and analysis

#### 3.1 Comparison with previous state-of-the-art methods.

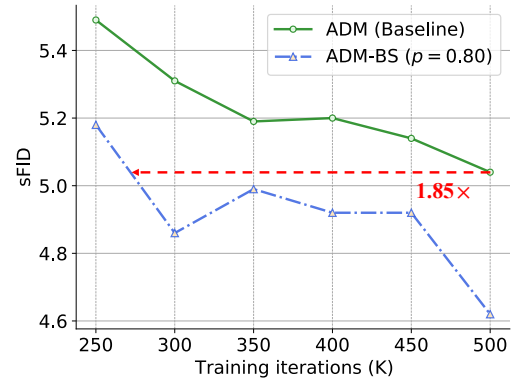
Table 1 in the manuscript compares the best performance of different methods. In this section, we compare the training acceleration effect of each previous state-of-the-art method in Figure 1, from which we can see that various re-weight methods accelerate the training speed and improve the generation quality, but our method achieves better acceleration effect and best generation quality. This further underscores the effectiveness of our ADM-BS approach.



**Figure 1: FID scores concerning the number of training iterations on AFHQ-D. All experimental configurations remain consistent with the previous settings.**



**Figure 2: sFID scores concerning the number of training iterations on CIFAR-10 dataset.**



**Figure 3: sFID scores concerning the number of training iterations on CelebA dataset.**

#### 3.2 More Quantitative Comparison.

**sFID.** sFID [14] is also a widely used metric to evaluate the sample quality. We compare our ADM-BS and ADM of this metric on CIFAR and CelebA datasets. The results, presented in Figure 2 and Figure 3, show that ADM-BS achieves a 2× faster attainment of the sFID score on CIFAR and a 1.85× faster rate on CelebA. Also, we note that ADM-BS consistently outperforms ADM in sFID on both datasets. **Precision and Recall.** We conduct a comparison using both Improved Precision and Recall metrics [11] to evaluate sample fidelity and diversity separately. Sample fidelity is quantified as the fraction of model samples that fall within the data manifold (precision), while diversity is measured as the fraction of data samples that fall within the sample manifold (recall). The result in Table 1 indicates that ADM-BS gets better results than ADM on all datasets. This observation suggests that the data generated by the ADM-BS model exhibits higher quality and greater alignment with the real data. Meanwhile, ADM-BS is also more proficient in capturing and reproducing a broader range of features or samples present in the training data. This illustrates the effectiveness of our Bernoulli Distribution-Based Sampling method in enhancing the generative quality of the model.

**Table 1: Quantitative Comparison. Comparing Precision and Recall of ADM and ADM-BS on the five datasets.**

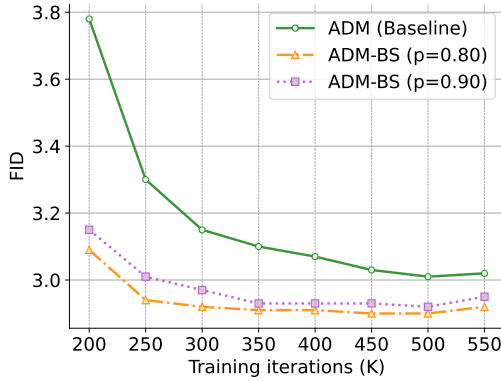
Model	CIFAR 32×32		CelebA 64×64		FFHQ 128×128		Stanford Cars 128×128		ADHQ-D 256×256	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
ADM	0.682	0.600	0.703	0.534	0.702	0.444	0.732	0.529	0.772	0.520
<b>ADM-BS</b>	<b>0.689</b>	<b>0.605</b>	<b>0.726</b>	<b>0.547</b>	<b>0.731</b>	<b>0.485</b>	<b>0.759</b>	<b>0.546</b>	<b>0.799</b>	<b>0.573</b>

**Table 2: Quantitative Comparison. Comparing the LPIPS for ADM and ADM-BS on the seven datasets.**

LPIPS ↓	CIFAR	ImageNet	CelebA	Stanford Cars	FFHQ	CelebAHQ	AFHQ-D
ADM	0.454	0.462	0.431	0.860	0.816	0.793	0.736
<b>ADM-BS</b>	<b>0.312</b>	<b>0.324</b>	<b>0.376</b>	<b>0.643</b>	<b>0.674</b>	<b>0.606</b>	<b>0.653</b>

**LPIPS.** We compare the Learned Perceptual Image Patch Similarity distances (LPIPS) [21] metric on different four datasets. The LPIPS distances align closely with human perception and are extensively used for evaluating generative models, offering an effective measure that mirrors human visual assessment. In our experiments, we averaged the LPIPS distances measured with 10K random samples. The result is shown in Table 2, we notice that ADM-BS outperforms the ADM significantly in all four datasets, demonstrating the effectiveness of ADM-BS.

**More discussion about  $p$ .** we also try combinations of  $p = 0.9$  and  $p = 1$  in the CIFAR-10 datasets. The results in Figure 4 indicates that  $p = 0.9$  yields performance better than the baseline but not as good as  $p = 0.8$ . However,  $p = 1$  performs very poor (FID=7.96). This indicates that the diffusion model need to be trained on timesteps greater than  $0.8T$ . In the inference stage, we start from Gaussian distribution. Although the different distributions become similar after  $0.8T$ , this stage still differs from the initial sampling point. Therefore, this timestep stage of training should not be overlooked.

**Figure 4: FID scores concerning the number of training iterations on ADM-IP and ADM-IP-BS on FFHQ dataset. All settings follow ADM-IP.**

### 3.3 Integrating with input perturbation method.

Recently, ADM-IP [16] finds that the long sampling chain of the diffusion model leads to an error accumulation bias. To reduce

this bias, ADM-IP proposes an input perturbation method in the training stage of the diffusion model. Since this method also uses uniform timestep sampling in the training stage, we combine our method with ADM-IP on CIFAR-10 and FFHQ datasets. The results on in Table 3 and Table 4 indicate that our method significantly improves the performance of ADM-IP. Meanwhile, we also compare the training iterations and FID scores on Figure 5 and Figure 6. ADM-IP-BS also enhances training speed by a factor of  $1.6 \times$  on CIFAR-10 and  $2.78 \times$  on FFHQ. This further illustrates the generalizability of our non-uniform timestep sampling methods.

**Table 3: Combine with ADM-IP. Each FID and sFID score is computed using  $T' = 100$  inference steps on CIFAR-10.**

Method	ADM	ADM-BS	ADM-IP	ADM-IP-BS
FID	3.47	3.30	2.70	<b>2.35</b>
sFID	5.73	4.80	4.51	<b>3.99</b>

**Table 4: Combine with ADM-IP. Each FID and sFID score is computed using  $T' = 100$  inference steps on FFHQ.**

Method	ADM	ADM-BS	ADM-IP	ADM-IP-BS
FID	14.07	10.73	7.07	<b>6.24</b>
sFID	13.14	11.75	8.35	<b>7.98</b>

### 3.4 Other distribution-based method.

In addition to the Bernoulli distribution-based time sampling, another possible timestep sampling method is based on the exponential distribution. Since the upper bound of the Wasserstein distance between different initial distributions is exponentially decreasing, we change the uniform timestep sampling to time sampling based on the exponential distribution (i.e., ADM-E), and the results of the experiments are shown in Figure 7. It can be seen that better FID and training speedups are achieved by ADM-E compared to ADM, but the performance is not as good as that of ADM-BS. However, this still shows that designing non-uniform timestep sampling methods focusing on significant intervals can be beneficial for the acceleration of diffusion model training and the improvement of generation quality.

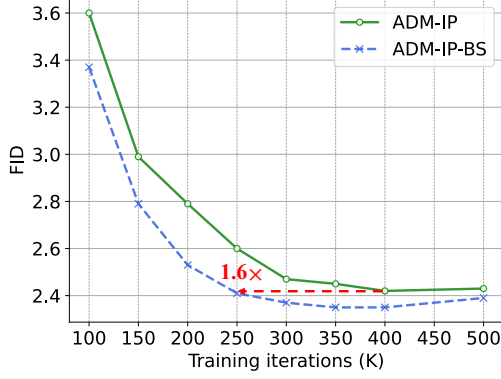


Figure 5: FID scores concerning the number of training iterations on ADM-IP and ADM-IP-BS on CIFAR-10 dataset. All settings follow ADM-IP.

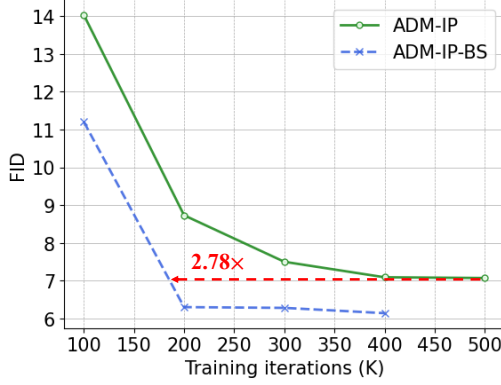


Figure 6: FID scores concerning the number of training iterations on ADM-IP and ADM-IP-BS on FFHQ dataset. All settings follow ADM-IP.

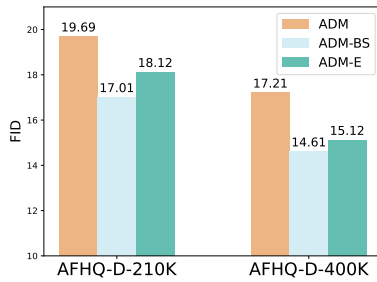


Figure 7: Comparing FID of different time sampling methods on AFHQ-D in various training iterations.

### 3.5 More Diffusion-based Model.

We also conduct large-scale experiments using the DiT-S/8 architecture [17], a prevalent multimodal backbone, with the ImageNet-256 dataset. The results in Figure 8 demonstrate that our non-uniform

timestep sampling method also benefits the DiT architecture, underscoring its importance in the fundamental diffusion model.

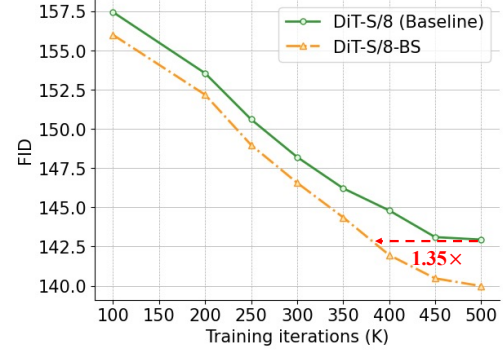


Figure 8: FID scores concerning the number of training iterations on ImageNet.

To further verify the generalizability of ADM-BS across different diffusion framework, we performed additional comparative analyses on the AFHQ-D dataset. In these experiments, we employ the log-normal noise distribution as proposed by EDM [8]. The results in Table 5 demonstrate that our non-uniform timestep sampling method is also beneficial to EDM noise distribution.

Table 5: More comparisons of different noise distribution.

FID	ADM	ADM-BS
Linear Schedule	17.21	14.08
Log-normal Schedule	15.42	12.56

## 4 Hyper-parameter

In this section, we provide detailed information about the architecture and training parameters of our ADM-BS method, as outlined in Table 6. During the training stage, we employ the AdamW optimizer, setting the hyperparameters ( $\beta_1$ ;  $\beta_2$ ) to (0.9; 0.999) across all models. Given that our method is compatible with existing inference techniques, we maintain standard parameter settings during the inference stage, consistent with those described in previous methods [3, 13, 15, 19].

## 5 Visual Results on Different Datasets

In this section, we present additional results generated using our ADM-BS method on different datasets. Figure 9 displays samples generated on the CelebA dataset [12], while Figures 10, 11, 12 and 13 show the results on the Stanford Cars [10], FFHQ [9], CelebAHQ [7] and AFHQ-D [2] datasets, respectively. For each dataset, we use the best number of training iteration and inference steps as indicated in the manuscript.

## References

- [1] Patrick Billingsley. 2017. *Probability and measure*. John Wiley & Sons.

**Table 6: The hyper-parameter of our ADM-BS method in different datasets.**

	CIFAR-10, ImageNet (32 × 32)	CelebA (64 × 64)	Stanford Cars, FFHQ (128 × 128)	AFHQ-D, CelebAHQ (256 × 256)
Diffusion Step	1000	1000	1000	1000
Noise Schedule	cosine	cosine	cosine	linear
Channels	128	192	256	128
Residual Blocks	3	3	3	1
Channel multiple	(1, 2, 2, 2)	(1, 2, 2, 2)	(1, 2, 2, 2)	(1, 1, 2, 2, 4)
Head Channels	32	64	64	64
Attention resolutions	(16, 8)	(32, 16, 8)	(32, 16, 8)	(16)
Learning Rate	1e-4	1e-4	1e-4	2e-5

- [2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *CVPR*. Computer Vision Foundation / IEEE, 8185–8194.
- [3] Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*. 8780–8794.
- [4] Sandesh Ghimire, Jinyang Liu, Armand Comas Massague, Davin Hill, Aria Ma-soomi, Octavia I. Camps, and Jennifer G. Dy. 2023. Geometry of Score Based Generative Models. *CoRR* abs/2302.04411 (2023).
- [5] Thomas Hakon Gronwall. 1919. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics* (1919), 292–296.
- [6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*.
- [7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *CoRR* abs/1710.10196 (2017).
- [8] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *NeurIPS*.
- [9] Tero Karras, Samuli Laine, and Timo Aila. 2021. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 12 (2021), 4217–4228.
- [10] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3D Object Representations for Fine-Grained Categorization. In *ICCVW*. IEEE Computer Society, 554–561.
- [11] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved Precision and Recall Metric for Assessing Generative Models. In *NeurIPS*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 3929–3938.
- [12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*. IEEE Computer Society, 3730–3738.
- [13] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. *arXiv preprint arXiv:2206.00927* (2022).
- [14] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W. Battaglia. 2021. Generating images with sparse representations. In *ICML (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 7958–7968.
- [15] Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. 2023. Elucidating the Exposure Bias in Diffusion Models. *arXiv preprint arXiv:2308.15321* (2023).
- [16] Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. 2023. Input Perturbation Reduces Exposure Bias in Diffusion Models. In *ICML (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 26245–26265.
- [17] William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. In *ICCV*. 4172–4182.
- [18] Hannes Risken and Hannes Risken. 1996. *Fokker-planck equation*. Springer.
- [19] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net.
- [20] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*. OpenReview.net.
- [21] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 586–595.





Figure 9: More visual results on CelebA datasets.

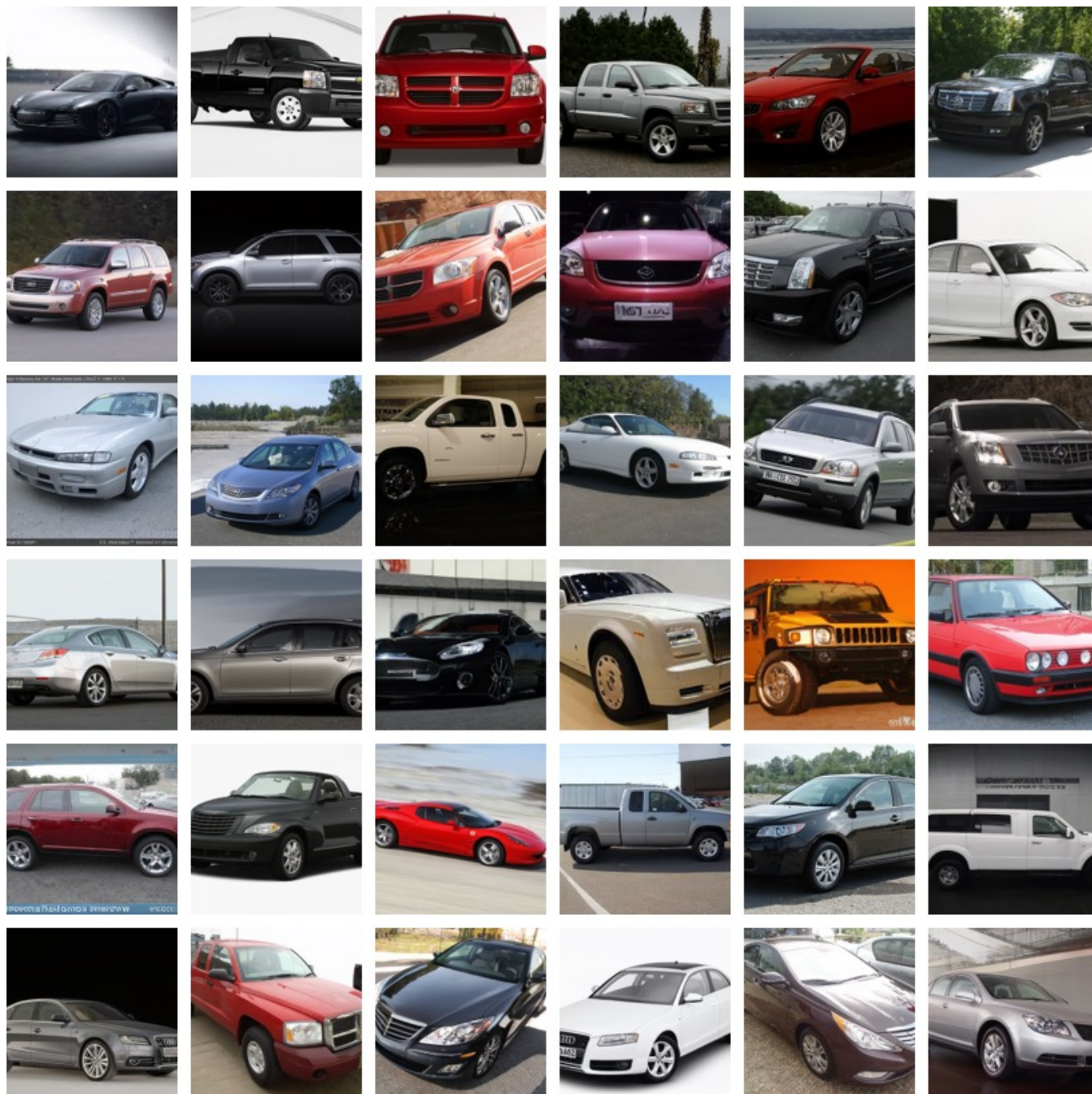


Figure 10: More visual results on Stanford Cars datasets.





Figure 11: More visual results on FFHQ datasets.





Figure 12: More visual results on CelebA HQ datasets.





Figure 13: More visual results on AFHQ-D datasets.