Beta-Tuned Timestep Diffusion Model Supplementary Material

Tianyi Zheng[1,2], Peng-Tao Jiang[2], Ben Wan[1], Hao Zhang[2], Jinwei Chen[2], Jia Wang[1] ✉, and Bo Li[2] ✉

[1] Shanghai Jiao Tong University, Shanghai, China
{tyzheng, burn-w, jiawang}@sjtu.edu.cn
[2] vivo Mobile Communication Co., Ltd, China
{pt.jiang,haozhang,jinwei.chen,libra}@vivo.com

1 Overview

In the appendix, we first present a detailed proof of the Lemma, Theorem, and Proposition in Section 2. In Section 3, we present additional experimental results and a detailed explanation of the experimental setups. Finally, we present more visual results in Section 4.

2 The detailed proof.

2.1 Preliminary

In this section, we first review the forward process and assumptions of DDPM. For any time t, the conditional distribution $q_t(x_t|x_0)$ in DDPM [2,3,5] is

$$q_t\left(x_t \mid x_0\right) = \mathcal{N}\left(x_t \mid e^{-\frac{t}{2}}x_0, (1-e^{-t})\mathbf{I}\right) \tag{1}$$

Meanwhile, we have the following assumption.

Assumption 1 *For the diffusion process described by Equation* (1)*, assume there exists a constant* $\delta > 0$*. For any* $t \in [0, T]$ *and any point* x_t *in high-density regions, the score* $\|\nabla_x \log q_t(x_t)\|$ *is bounded by* δ*, i.e.,*

$$\|\nabla_x \log q_t(x_t)\| \leq \delta.$$

This ensures that the score of the data distribution is bounded in the vicinity of high-density data throughout the diffusion process.

✉ Corresponding authors.
This work was done during Tianyi Zheng's internship at vivo.

2.2 Proof for Lemma 1

Lemma 1. *Consider the forward diffusion process as described by Equation (1). When $t \to 0^+$, the time derivative of the distribution $q_t(x_t)$ is predominantly governed by the second-order derivative term, encapsulated in the relationship:*

$$\frac{\partial}{\partial t} q_t(x_t) = \nabla_{x_t}^2 q(x_t), \tag{2}$$

Proof of Lemma 1. Consider the forward diffusion process as described by the probability density function $q_t(x_t)$, given by

$$q_t(x_t) = \frac{1}{\sqrt{2\pi\left(1 - e^{-2t}\right)}} \exp\left(-\frac{\left(x_t - e^{-t}x_0\right)^2}{2\left(1 - e^{-2t}\right)}\right). \tag{3}$$

We aim to compute the time derivative $\frac{\partial}{\partial t} q_t(x_t)$ and show that as $t \to 0^+$, it is predominantly influenced by the second-order spatial derivative $\nabla_{x_t}^2 q_t(x_t)$. Based on the Fokker-Planck Equation of Equation (1), we have the following relationship

$$\frac{\partial}{\partial t} q(x_t) = \nabla_{x_t} q(x_t) + \nabla_{x_t}^2 q(x_t). \tag{4}$$

First, we compute the first-order spatial derivative with respect to x_t, denoted by $\nabla_{x_t} q_t(x_t)$. Applying the chain rule, we have

$$\nabla_{x_t} q_t(x_t) = -\frac{(x_t - e^{-t}x_0) \exp\left(-\frac{(x_t - e^{-t}x_0)^2}{2(1-e^{-2t})}\right)}{\sqrt{2\pi(1-e^{-2t})}(1-e^{-2t})}. \tag{5}$$

Next, we find the second-order spatial derivative $\nabla_{x_t}^2 q_t(x_t)$ by differentiating $\nabla_{x_t} q_t(x_t)$ with respect to x_t again:

$$\nabla_{x_t}^2 q_t(x_t) = \frac{\exp\left(-\frac{(x_t - e^{-t}x_0)^2}{2(1-e^{-2t})}\right)}{\sqrt{2\pi(1-e^{-2t})}} \left[-\frac{1}{(1-e^{-2t})} + \frac{(x_t - e^{-t}x_0)^2}{(1-e^{-2t})^2}\right]. \tag{6}$$

As $t \to 0^+$, the term e^{-2t} approaches 1 from below, causing the denominator $1 - e^{-2t}$ to approach 0, leading to a rapid increase in the magnitude of the second-order term compared to the first-order term. This indicates that the second-order spatial derivative term $\nabla_{x_t}^2 q_t(x_t)$ becomes the dominant factor in the time derivative $\frac{\partial}{\partial t} q_t(x_t)$, as postulated. □

2.3 Proof for Theorem 2

Theorem 2. *Consider the forward diffusion process described by Equation (2). The norm of the time derivative of the distribution $q_t(x_t)$ is bounded by the following inequality:*

$$\left\| \frac{\partial}{\partial t} q_t(x_t) \right\| \leq \left\| \frac{q_t(x_t)}{2t} \right\|. \tag{7}$$

As $t \to 0^+$, the norm of the derivative of the distribution $q_t(x_t)$ may become unbounded. This is indicated by the limit superior:

$$\limsup_{t \to 0^+} \left\| \frac{q_t(x_t)}{2t} \right\| = \infty. \tag{8}$$

Proof of Theorem 2. Consider the $H(t,x) = \frac{1}{(4\pi t)^{\frac{n}{2}}} e^{-\frac{|x|^2}{4t}}$, which is a fundamental solution to the Equation (2) and hence satisfies Equation (2). We have the following relationship of $H(t,x)$:

$$\nabla_x \log H(t,x) - \frac{\partial_t H}{H} = \frac{x}{2t}. \tag{9}$$

Given a non-negative, continuous, bounded function $q_0(x_0)$, the solution $q_t(x_t)$ at time t can be expressed as a convolution with the $H(t,x)$:

$$q_t(x_t) = (H(t,\cdot) * q_0)(x_t). \tag{10}$$

Define $\nu_{t,x}(y) = \frac{H(t,x-y)q_0(y)}{q_t(x_t)}$, we can verify that $\int_{\mathbb{R}^n} \nu_{t,x}(y)\, dy = 1$, making $\nu_{t,x}(y)$ a probability measure of y. Then we have the following relationship:

$$\nabla q_t(x_t) = \int_{\mathbb{R}^n} \nabla H(t, x-y) q_0(y)\, dy, \tag{11}$$

$$\partial_t q_t(x_t) = \int_{\mathbb{R}^n} \partial_t H(t, x-y) q_0(y)\, dy. \tag{12}$$

Utilizing the relationship established for $H(t,x)$, we derive the following inequality for the score of the $q_t(x_t)$:

$$\nabla_x \log q_t(x_t) = \left| \int_{\mathbb{R}^n} \frac{\nabla H(t,x-y)}{H(t,x-y)} \nu_{t,x}(y)\, dy \right|^2 \tag{13}$$

$$\leq \int_{\mathbb{R}^n} \left| \frac{\nabla H(t,x-y)}{H(t,x-y)} \right|^2 \nu_{t,x}(y)\, dy \tag{14}$$

$$= \int_{\mathbb{R}^n} \left(\frac{\partial_t H(t,x-y)}{H(t,x-y)} + \frac{n}{2t} \right) \nu_{t,x}(y)\, dy \tag{15}$$

$$= \frac{\partial_t q_t(x_t)}{q_t(x_t)} + \frac{n}{2t}. \tag{16}$$

Given that $\partial_t q_t(x_t) < 0$ in the forward diffusion process, and considering Assumption 1, we conclude with the following bound on the time derivative of $q_t(x_t)$:

$$\left\| \frac{\partial}{\partial t} q_t(x_t) \right\| \leq \left\| \frac{q_t(x_t)}{2t} - \delta q_t(x_t) \right\| \leq \left\| \frac{q_t(x_t)}{2t} \right\|. \tag{17}$$

□

2.4 Proof for Proposition 2

Proposition 2. *Given a timestep sampling method utilizing the B-TTDM with a Beta distribution parameterized by $\alpha < 1$ and $\beta = 1$, the tuned error at the initial time position t is $\left(\frac{1}{T}\right)^{\alpha}$, where T is the total diffusion steps.*

Proof of Proposition 2. First, we recall the probability density function (pdf) of the Beta distribution, given by

$$f(t;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1}, \tag{18}$$

where $\Gamma(\cdot)$ denotes the Gamma function. Setting $\beta = 1$, the pdf simplifies to

$$f(t;\alpha,1) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(1)} t^{\alpha-1} = \alpha t^{\alpha-1}, \tag{19}$$

where we use $\Gamma(1) = 1$ and $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.

The cumulative probability of selecting a timestep within the interval $[0, \frac{1}{T}]$ is given by the integral of $f(t;\alpha,1)$ over this interval:

$$\int_0^{\frac{1}{T}} f(t;\alpha,1)\, dt = \int_0^{\frac{1}{T}} \alpha t^{\alpha-1}\, dt = \left(\frac{1}{T}\right)^{\alpha}. \tag{20}$$

Hence, the tuned error of selecting a timestep within the initial interval $[0, \frac{1}{T}]$ under the B-TTDM is $\left(\frac{1}{T}\right)^{\alpha}$.

□

3 Additional results and analysis

3.1 More Quantitative and Qualitative Comparison.

Parameters of the beta distribution. Our analysis in manuscript Section 3.3 highlight that the $\alpha = 1, \beta > 1$ configurations don't fully align with the forward diffusion's non-uniform properties. In this section, We try this α and β configurations on AFHQ-D datasets. The results in Table 1 show that this configurations also improves the performance of the diffusion models, but does not reach our improvement. This emphasizes the limitations of uniform timestep sampling in the training stage of diffusion model.

More Diffusion Framework. To examine B-TTDM's generalizability across different diffusion models, we conduct further experiments within various diffusion model frameworks. We comapre B-TTDM with NCSN [7] and EDM [4] on the CIFAR-10 dataset. With B-TTDM, NCSN's FID improved from **11.89** to **10.24**, despite its simpler U-Net architecture. Moreover, EDM uses a log-normal distribution and set greater weight to "middle" noise values. When adjusted to follow the B-TTDM trend, the FID score improves from **1.92** to **1.81**. Therefore, B-TTDM has greater generalizability and is suitable for a wider range of diffusion frameworks.

Table 1: Ablation study on the parameters of Beta Distribution for AFHQ-D (256×256).

Beta Distribution		AFHQ-D (256×256)			
α	β	FID	sFID	Recall	Precision
1	1.2	15.44	48.82	0.570	0.778
1	**1.5**	**15.00**	**45.43**	**0.629**	0.715
1	1.8	16.46	48.70	0.561	**0.798**
1	2	**1**7.12	50.52	0.553	0.766
1	1	17.21	49.03	0.553	0.738

Latent Diffusion Model. Recently, the DiT [6] has been widely adopted in latent diffusion model. The DiT achieves impressive generation results by compressing the image into a latent space using VAE before diffusion. In Figure 1, we compare the performance of B-TTDM within the **lightweight DiT-S/8** in latent space on ImageNet256 dataset. The results demonstrate that B-TTDM also enhances the performance of the latent diffusion model.

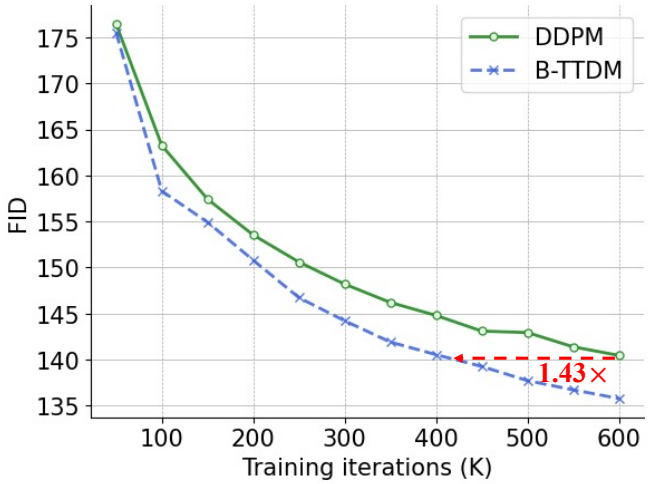

Fig. 1: FID scores concerning the number of training iterations on ImageNet (256×256).

More Qualitative Comparison. We present a visual comparison between B-TTDM and other re-weighting methods in Figure 2 on AFHQ-D and CelebA-HQ. It is evident from the figure that the images generated by B-TTDM exhibit better quality details. We randomly choose the three generated images without cherry-pick in each dataset.

Fig. 2: Qualitative comparison. Unconditional generation results for B-TTDM and other re-weighting methods

3.2 Hyper-parameter

In this section, we list more details about the hyper-parameter and training details of B-TTDM in Table 2. It is worth noting that for large resolution (256 × 256) datasets, we use a lighter version of ADM [2] following previous work [1].

Table 2: The hyper-parameter of B-TTDM method in different resolution datasets.

	32 × 32	64 × 64	128 × 128	256 × 256
Diffusion Step	1000	1000	1000	1000
Noise Schedule	Cosine	Cosine	Cosine	Linear
Channels	128	192	256	256
Residual Blocks	3	3	3	3
Channel multiple	(1, 2, 2, 2)	(1, 2, 2, 2)	(1, 2, 2, 2)	(1, 1, 2, 2, 4)
Head Channesl	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

4 Visual Results on Different Datasets

In this section, we present additional results of unconditional generation using our B-TTDM method across various datasets. We utilize the B-TTDM model as reported in Figure 3 of the manuscript.

Fig. 3: More visual results on CIFAR 32 × 32. (FID = 2.89, 1000 inference steps.)

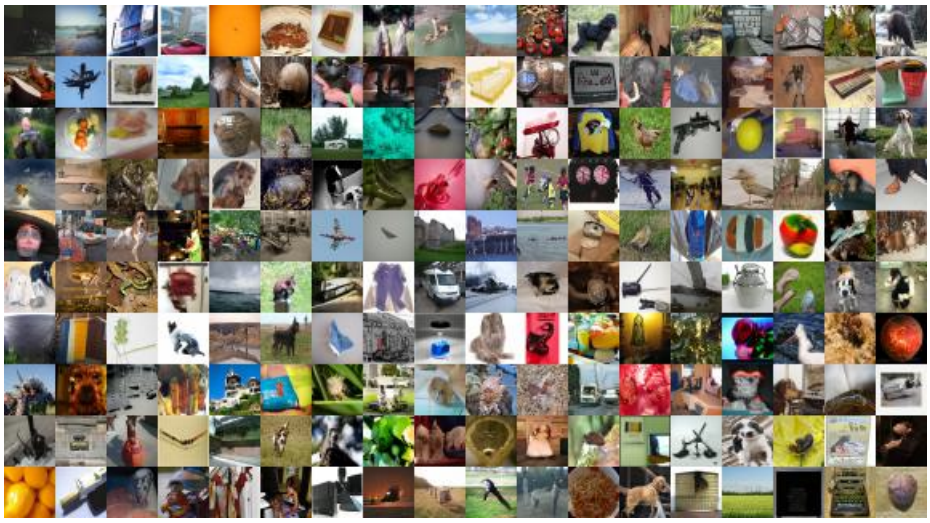

Fig. 4: More visual results on ImageNet 32 × 32. (FID = 2.87, 1000 inference steps.)

Fig. 5: More visual results on Celeba 64 × 64. (FID = 2.98, 300 inference steps.)

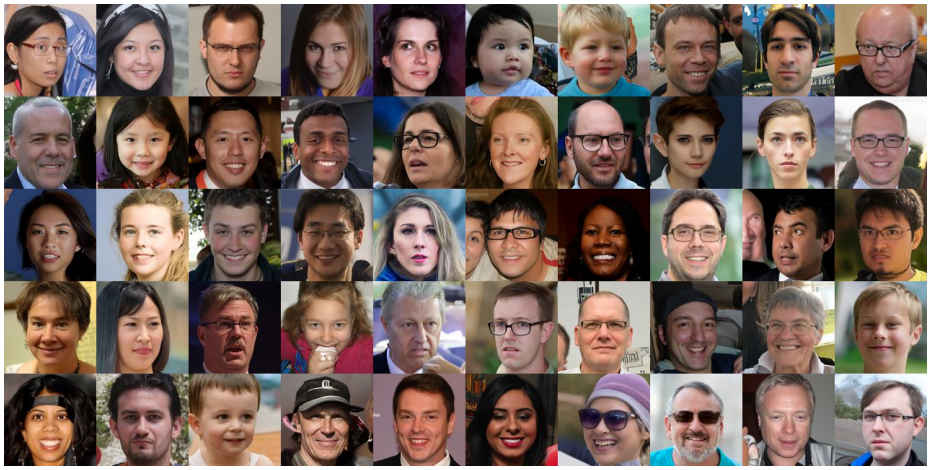

Fig. 6: More visual results on FFHQ 128 × 128. (FID = 10.28, 100 inference steps.)

Fig. 7: More visual results on CelebAHQ 256 × 256. (FID = 13.62, 100 inference steps.)

Fig. 8: More visual results on AFHQ-D 256 × 256. (FID = 13.81, 100 inference steps.)

References

1. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception prioritized training of diffusion models. In: CVPR. pp. 11462–11471. IEEE (2022)
2. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. In: NeurIPS. pp. 8780–8794 (2021)
3. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
4. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: NeurIPS (2022)
5. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8162–8171. PMLR (2021)
6. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4172–4182 (2023)
7. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. In: NuerIPS (2020)