MBSRapid: Momentum-Based Score Reutilization for Accelerated Sampling in Diffusion Models

Guangyi Wang [1], **Yong Chen**[2], **Hongpeng Chen**[2]
[1]School of Informatics, Xiamen University
[2]Institute of Artificial Intelligence, Xiamen University
{31520231154316, 36920231153185, 36920231153183}@stu.xmu.edu.cn

Abstract

Since their rise in 2020, Diffusion Probability Models (DPMs) have achieved revolutionary breakthroughs in the field of image generation, thanks to their straightforward optimization processes. However, compared to one-step generative paradigms such as Generative Adversarial Networks (GANs), the multi-time-step iterative sampling process of DPMs leads to a significant decrease in sampling efficiency. In this paper, we draw on methods for optimizing Stochastic Gradient Descent (SGD) and introduce a momentum-based score reutilization sampler designed to accelerate the sampling process in diffusion models. Notably, this method is train-free, can be directly applied to pre-trained diffusion models and is orthogonal to existing accelerated sampling algorithms. More specifically, our proposed sampler reuses the score predicted by the network through a momentum mechanism, integrating score estimates at different scales, thereby enabling samples to converge more quickly and smoothly to the target distribution. The experimental results indicate that our method reduced the FID to 3.18 on the CIFAR10 dataset, which is a 36.1% improvement over the DDPM baseline of 4.98. Additionally, during the sampling process with fewer steps, our sampling speed was twice that of the baseline.

Introduction

Since their introduction, deep learning generative models like Variational Autoencoders Bayes (VAEs) (Kingma and Welling 2013) in 2014 and Generative Adversarial Networks (GANs) (Creswell et al. 2018) in 2015 have significantly impacted various fields such as image generation, 3D reconstruction, and speech generation. These models aim to generate samples that mimic a training data distribution, with $x \sim p_\theta(x \mid z)$ representing the model's output given a sample z from a prior distribution $z \sim p(z)$. GANs have faced challenges like mode collapse and limited diversity, while VAEs often produce blurry images due to an imperfect surrogate objective ($ELBO$). However, the emergence of Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) in 2020 marked a significant advancement. Diffusion models, focusing on minimizing the L2 norm between outputs and noise, have simplified the optimization process and improved sample quality, leading them to become the current research focus in deep generative models (Croitoru et al. 2023; Yang et al. 2023).

However, DDPMs employ a thousand-step Markov chain process to infuse images with noise, diffusing them into a standard Gaussian noise state, subsequently leveraging neural networks to learn the noise distribution and reverse the trajectory for image restoration. Notably, this Markov chain necessitates a thousand neural network invocations, significantly impeding the model's sampling efficiency. Therefore, compared to one-step generative models like GANs, sampling efficiency remains a primary shortfall for DPMs (Zhou et al. 2023; Wang et al. 2023; Zheng et al. 2023a). In response to this challenge, extensive research efforts have been dedicated to enhancing DDPMs' sampling efficiency. Noteworthy advancements include: accelerated sampling methods that, by adjusting the sampling formula, negate the need for additional training, such as DDIM (Song, Meng, and Ermon 2020), and DPM-Solver series methods (Lu et al. 2022a,b; Zheng et al. 2023b); strategies based on knowledge distillation, like Consistency Models (Song et al. 2023) and Rectified Flow (Liu, Gong, and Liu 2022); and various accelerated sampling algorithms, for instance, AutoDiffusion (Li et al. 2023), which achieves training-free sampling acceleration by optimizing the time steps and network architecture through search algorithms.

The acceleration algorithms based on knowledge distillation mentioned above are capable of compressing Markov chains with thousands of steps into one step. However, such methods often require a considerable amount of additional training resources. Meanwhile, other training-free acceleration algorithms, represented by the DPM-Solver series, have made up-to-date progress in reducing the number of sampling steps to under 10 steps with only a slight compromise in sample quality. Yet, there is still a notable gap when compared to one-step sampling algorithms exemplified by GANs. Therefore, it is imperative to further reduce the number of sampling steps without sacrificing quality in training-free acceleration algorithms. Inspired by SGD and Langevin dynamics sampling methods, this paper proposes an innovative momentum-based score reutilization sampling strategy, orthogonal to existing acceleration algorithms and train-free. It enhances sampling efficiency on top of the current training-free acceleration methods without compromising quality. Specifically, since the parameters for the predicted

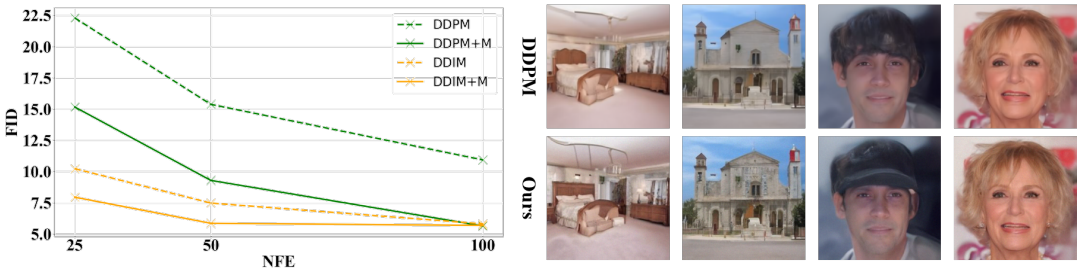

Figure 1: Performance Demonstration of MBSRapid Accelerated Sampling. Left: Performance comparison between our proposed MBSRapid and the DDPM and DDIM methods on the CIFAR10 dataset with different NFE settings (25, 50, 100). Right: Comparative results on the high-resolution LSUN and CelebA-HQ datasets with NFE set to 25 and momentum term m set to 0.15, between DDPM and our proposed method.

noise distribution at each step are shared by the same neural network, the scores at different scales inherently reflect the information of the same training data distribution. Accordingly, our proposed method introduces a momentum mechanism, considering the correlations between different steps in the iterative process, and allows for the reutilization of previously computed scores. Such reutilization not only increases computational efficiency but also reduces the loss of inferred sample information, thus fostering faster and more stable convergence of the samples to the target distribution.

Our experimental results demonstrate that the proposed method can maintain or even enhance the quality of sampling with significantly reduced steps, paving a new way for accelerated sampling in diffusion models. Overall, this paper highlights the current deficiencies in the sampling efficiency of widely-focused DPMs and introduces a momentum-based score reutilization strategy for accelerated sampling that requires no additional training.

Related Work

Accelerated Sampling Based on Sampler Design

DDIM (Song, Meng, and Ermon 2020) marks a significant breakthrough in accelerated sampling research by deconstructing the limitations of Markov chains and eliminating the dependency on $p(x_t|x_{t-1})$, thereby accelerating the sampling process. DDIM reassesses the inherent Markov assumption within DDPMs, noting that the relevance of the current state is not only with the preceding state but also pertains to earlier ones. This insight has prompted the theoretical derivation of a new sampling algorithm that doesn't require a complete time step sequence to achieve the sampling goal. Notably, DDIM enables the execution of sampling using any subset of the original time steps, achieving rapid sampling in no more than 100 steps and garnering widespread attention in academia.

Currently, the DPM-Solver series (Lu et al. 2022a,b; Zheng et al. 2023b) represents the state-of-the-art in accelerated sampling. It relies on a unified framework based on ScoreSDE, using the semi-linear structure of DPMs to deduce analytical solutions from an ODE perspective and employs alternative techniques for approximate calculations of these solutions. Through three iterative versions, DPM-Solver accomplishes DPMs sampling within 10 steps without additional training, standing as the fastest method available. Furthermore, DPM-Solver reveals that DDIM is essentially its first-order method, and its higher-order methods can achieve smaller error rates, hence surpassing DDIM's sampling capabilities within 10 steps.

Methods Based on Knowledge Distillation

Consistency Models (Song et al. 2023) and Rectified Flow (Liu, Gong, and Liu 2022) achieve "one-step sampling" by adopting knowledge distillation strategies. The consistency model, by modeling $f(x_t, t) = f(x_{t'}, t')$, compresses the diverse outputs of DPMs at each time step into uniform results, allowing each time step to directly produce the sampling target x_0. Concurrently, Rectified Flow models the sampling trajectory of DPMs as a linear path and distills a $v(x_t, t)$ model through training. By exploiting this "velocity" model and the characteristics of a linear trajectory, one-step sampling is readily achieved. Although distillation-based methods facilitate efficient sampling, they typically require a substantial investment in training resources.

Other Accelerated Sampling Strategies

AutoDiffusion (Li et al. 2023) takes off from the pre-trained model structure of DPMs and the redundancy of time steps, significantly reducing the computational burden of the sampling process by directly simplifying time steps and model structures through optimization search algorithms. Notably, this method has been demonstrated to be orthogonal to other accelerated sampling techniques, potentially enhancing the speed of existing efficient sampling methods, such as DPM-Solver, even further.

Method

Motivation

Inspired by the concept of SGD (Robbins and Monro 1951) and the Langevin dynamics sampling algorithm (Welling and Teh 2011) used in NCSNs (Song and Ermon 2019), as follows:

$$x_{i+1} = x_i + \epsilon \nabla_x \log p(x) + \sqrt{2\epsilon} Z_i, i = 0, 1, \ldots, K, \quad (1)$$

where $z_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), x_0 \sim \pi(x)(\pi(x)$ is any arbitrary prior distribution); α represents the step size. When $\alpha \to 0, K \to$

∞, x_K converges to the distribution $p(x)$. we observe that the form of Equation (1) is similar to that of the SGD optimization algorithm. The $\sqrt{2\epsilon}Z_i$ term in it introduces necessary randomness but does not alter the distribution itself. If we consider ϵ as the learning rate in SGD, then the loss function of SGD can be analogized to $-\log p(x)$. In this context, the goal of SGD is to minimize this loss function, which is equivalent to maximizing the likelihood function $\log p(x)$. Considering the mathematical similarity between Langevin dynamics sampling (i.e., DPMs sampling algorithm) and SGD, we propose using optimization techniques of SGD, such as momentum methods (Polyak 1964), RMSProp, Adam (Kingma and Ba 2014), etc., to accelerate or smooth the sampling process of DPMs. Specifically, we introduce a Langevin dynamics sampling algorithm with momentum mechanism as shown in Equation (2):

$$\begin{aligned} v_{i+1} &= mv_i - \epsilon \nabla_x \log p(x) + \sqrt{2\epsilon} Z_i, \\ x_{i+1} &= x_i - v_{i+1}, \end{aligned} \tag{2}$$

where m represents the momentum coefficient, with a value range between 0 and 1 . The variable v accumulates previous scores. By reutilizing these scores, we are able to integrate information across different time scales, thereby achieving faster and smoother convergence of samples to the target distribution. This method is orthogonal to other accelerated sampling algorithms.

Momentum-based Accelerated Sampler

In addition to the score-based DPMs such as NCSNs, there are also DPMs founded on variational theory, such as DDPMs and DDIM. Notably, the ScoreSDE (Song et al. 2020) research successfully unified these DPM perspectives using Stochastic Differential Equations (SDE), proposing a unified diffusion framework that employs neural networks to predict score $\nabla_x \log p(x)$, combined with SDE or Ordinary Differential Equation (ODE) solvers for sample generation. This implies that for all DPMs, whether from a variational or score perspective, the outputs predicted by neural networks can be directly or indirectly considered as score, thus likening the DPM sampling process to the SGD optimization algorithm. Hence, all DPMs can reuse score via the momentum mechanism without additional training burden, thereby accelerating sampling.

For all DPMs, we can unify the sampling methods as per Equation (3):

$$x_{t-1} = \lambda x_t - \eta dx_t + \xi z_t, t = 0, 1, \ldots, T, \tag{3}$$

where $z_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, with λ and ξ as coefficients for adjusting weights, η as the step size akin to a "learning rate", and dx_t as the neural network output at step t. For DDPMs, the sampling formula is expressed as Equation (4):

$$x_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t + \sigma_t z_t, \tag{4}$$

where $\alpha_t, \bar{\alpha}_t$, and x_0 are all defined by $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} z$, with x_0 being derived from x_t and $z_\theta\left(x_t, t\right)$. Additionally, σ_t^2 is equivalent to $\widetilde{\beta}_t = \left(\left(1-\bar{\alpha}_{t-1}\right) / \left(1-\bar{\alpha}_t\right)\right) \cdot \beta_t$ or β_t. We derive λ as $\left(\sqrt{\bar{\alpha}_{t-1}}\beta_t + \sqrt{\alpha_t \bar{\alpha}_t}\left(1-\bar{\alpha}_{t-1}\right)\right) / \left(1-\bar{\alpha}_t\right) \sqrt{\bar{\alpha}_t}$, η as $\left(1-\alpha_t\right) \sqrt{\bar{\alpha}_{t-1}} / \sqrt{\bar{\alpha}_t\left(1-\bar{\alpha}_t\right)}, dx_t$ as $z_\theta\left(x_t, t\right)$, and ξ as σ_t. For NCSNs, the primary sampling formula is the Langevin dynamics sampling algorithm, as shown in Equation (1), yielding λ as $1, \eta$ as $-\epsilon, dx_t$ as $\nabla_x \log p(x)$, and ξ as $\sqrt{2\epsilon}$. In DDIM, the sampling method is given by Equation (5):

$$\begin{aligned} x_{t-1} =& \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t} z_\theta\left(x_t, t\right)}{\sqrt{\bar{\alpha}_t}}\right) + \\ & \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} z_\theta\left(x_t, t\right) + \sigma_t z_t, \end{aligned} \tag{5}$$

in this formulation, $\sigma_t^2 = \tau^2 \widetilde{\beta}_t$. When τ is set to 1 , the model reverts to the traditional DDPM sampling form as described in Equation (4). Conversely, when τ is set to 0, it transitions to the sampling method of DDIM. Based on Equation (5), we can derive for the DDIM sampling algorithm that λ as $\sqrt{\bar{\alpha}_{t-1}} / \sqrt{\bar{\alpha}_t}, \eta$ as $\sqrt{\left(1-\bar{\alpha}_t\right) \bar{\alpha}_{t-1} / \bar{\alpha}_t}$ $\sqrt{1-\bar{\alpha}_{t-1}}, dx_t$ as $z_\theta\left(x_t, t\right)$, and ξ as 0 . Thus, we have unified the mentioned DPMs sampling algorithms into the form of Equation (3). Building upon this, by introducing a momentum mechanism, we arrive at a unified formula for momentum-based accelerated sampling as shown in Equation (6):

$$\begin{aligned} v_{t-1} &= mv_t + (1-m) \cdot \left(\eta dx_t - \xi z_t\right), \\ x_{t-1} &= \lambda x_t - v_{t-1}. \end{aligned} \tag{6}$$

Improved Momentum-Based Accelerated Sampler

The momentum mechanism, since its inception in deep learning, has undergone several iterations. For instance, the introduction of "predictive" updates in Nesterov Momentum (Nesterov 1983) has been proven more robust than traditional momentum methods in practical applications. In addition, the Adam method (Kingma and Ba 2014) can adjust momentum weights, effectively integrating features at different scales. Utilizing these advancements in momentum mechanisms, we can apply both Nesterov Momentum and Adam methods to the sampling process in DPMs, using score from various time scales to accelerate sampling. With Nesterov Momentum, we predict the score for the next time step to expedite convergence, as shown in Equation (7):

$$\begin{aligned} v_{t-1} &= mv_t + (1-m) \cdot \left(\eta d\left(\lambda x_t - mv_t\right) - \xi z_t\right) \\ x_{t-1} &= \lambda x_t - v_{t-1} \end{aligned} \tag{7}$$

Using the Adam method, we can adaptively integrate score from different time scales, thus accelerating the sampling process, as detailed in Equation (8):

$$\begin{aligned} v_{t-1} &= \left(\mu_1 v_t + (1-\mu_1) \cdot \left(\eta dx_t - \xi z_t\right)\right) / \left(1-\mu_1^t\right), \\ c_{t-1} &= \left(\mu_2 c_t + (1-\mu_2) \cdot \left(\eta dx_t - \xi z_t\right)^2\right) / \left(1-\mu_2^t\right), \\ x_{t-1} &= \lambda x_t - \frac{v_{t-1}}{\sqrt{c_{t-1}} + \zeta}, \end{aligned} \tag{8}$$

where v_{t-1} and c_{t-1} represent the first and second moment estimates of the score, respectively. The hyperparameters μ_1

and μ_2 are typically set to 0.9 and 0.999 , respectively. Additionally, ζ is a small constant used to prevent division by zero, set to 10^{-8} in this paper.

To conclude, we have thoroughly introduced the complete suite of improvements for our momentum-based score reutilization sampler, incorporating momentum, Nesterov Momentum, and Adam methods into our approach. This leads to a unified update formula: $x_{t-1} = \mathrm{MBSRapid}\,(x_t + \xi z_t, m, \mu_1, \mu_2)$, thus establishing our sampling algorithm as presented in Algorithm 1.

Algorithm 1: MBSRapid Algorithm

Input: $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
Parameter: m, μ_1, μ_2, ξ
Output: x_0

Initialization: $x_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), m, \mu_1, \mu_2, \xi$
for t $= \mathrm{T}, \ldots, 1$ **do**
    $z_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if t>1, else $z_t = \mathbf{0}$
    $x_{t-1} = \mathrm{MBSRapid}\,(x_t + \xi z_t, m, \mu_1, \mu_2)$
end for
return x_0

Experiments

Experimental Setup

To validate the effectiveness of our proposed method and its orthogonality to existing approaches, we selected four datasets for experimentation: CIFAR10 (Krizhevsky, Hinton et al. 2009) (32×32), CelebA (Liu et al. 2018) (32×32), CelebA-HQ (Karras et al. 2017) (256×256), and LSUN (Yu et al. 2015) (256×256). The performance metrics used in the experiments were FID and IS.

Regarding baseline models, considering that DDIM disrupts the Markov chain assumption in DDPM, any subset of the original sampling step sequence in DDPM and NCSN algorithms can complete the sampling process. Hence, we introduced our proposed sampler on top of the baselines of DDPM, DDIM, and NCSN algorithms. In Equation (4), DDPM's σ_t^2 takes two different values: β_t and $\widetilde{\beta}_t$. Studies have shown that β performs better in 1000-step sampling, while $\widetilde{\beta}_t$ is more effective in fewer-step sampling (Bao et al. 2022). Based on this, to verify the accelerated sampling effect of our algorithm, the $\widetilde{\beta}_t$ setting was used in the experiments. In fewer-step sampling experiments, we employed 25, 50, and 100 steps for Number of Function Evaluations (NFE) (Lu et al. 2022a) configurations. In the ablation experiments, we observed that the sole use of the momentum method outperforms Nesterov Momentum and Adam. Therefore, unless otherwise stated, the methods mentioned in this paper refer to the baseline model combined with the momentum strategy.

Overall Performance

In Table 1, we present a comprehensive comparison of the proposed method in this study with existing benchmarks such as DDPM, DDIM, and NCSN. Utilizing the CIFAR10 dataset and under the setting of 1000 NFEs, the results from

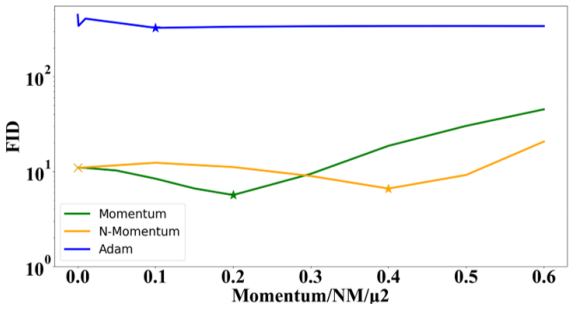

Figure 2: Optimal Momentum Strategy Search on the CIFAR10 Dataset Using the DDPM Approach with NFE Set to 100, involving Momentum, Nesterov Momentum, and Adam. The vertical axis represents the FID on a logarithmic scale, with the '×' marker indicating the baseline method, and the '*' marker denoting the optimal result for the current curve (same below).

Table 1 demonstrate a significant improvement in FID and IS metrics by our method over these three benchmarks. For instance, building upon the DDPM approach, our method reduced the FID score of the CIFAR10 dataset from 4.98 to 3.18, marking an enhancement of 36.1% compared to the baseline method. This outcome unequivocally indicates that our approach can generate samples of higher quality under the same NFE conditions.

In Figure 1, we further validate the effectiveness of our method in accelerated sampling. Observing Figure 1 (left), on the CIFAR10 dataset, we conducted comparative experiments using both DDPM and DDIM methods under the settings of 25, 50, and 100 NFEs. The results reveal that, after incorporating our proposed momentum mechanism (represented by solid lines), it consistently outperforms the equivalent baseline methods (depicted by dashed lines), as the solid lines are always positioned below and to the left of the dashed lines. This confirms the significant advantage of our method. Moreover, Table 2 lists the FID comparison results combining our method with the three benchmarks across all NFE settings. The data indicate performance enhancements in all baseline methods and NFE settings with our approach.

Sampler	CIFAR10	
	IS↑	FID↓
DDPM	9.06 ± 0.08	4.98
Ours	**9.41 ± 0.13**	**3.18**
DDIM	9.08 ± 0.09	4.28
Ours	**9.18 ± 0.08**	**3.54**
NCSN	8.83 ± 0.10	24.35
Ours	**8.98 ± 0.11**	**22.20**

Table 1: Comparative Results of Different Benchmark Methods and Our Method on the CIFAR10 Dataset with NFE Set to 1000 (FID score).

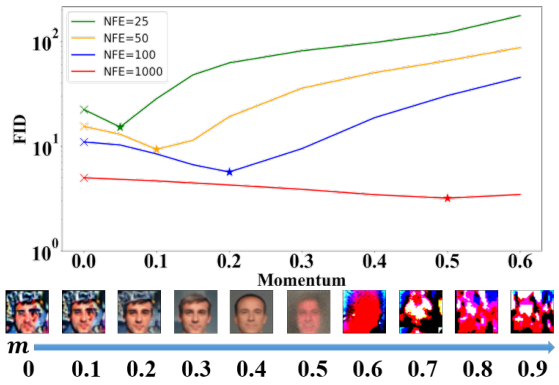

Figure 3: Optimal Search Range for the Momentum Hyperparameter m. Top: Performance comparison on the CIFAR10 dataset using the DDPM method across all NFE settings, by adjusting the value of the hyperparameter m. Bottom: Visual comparison of effects on the CelebA dataset using the NCSN method with NFE set to 100, by altering the value of the hyperparameter m.

Sampler	NFE			
	25	50	100	1000
DDPM	22.30	15.39	10.94	4.98
Ours	**15.17**	**9.30**	**5.65**	**3.18**
DDIM	10.24	7.48	5.82	4.28
Ours	**7.95**	**5.86**	**5.69**	**3.54**
NCSN	407.4	304.3	201.2	24.35
Ours	**407.2**	**304.2**	**201.1**	**22.20**

Table 2: Comparison Results of Different Benchmark Methods and Our Method Across All NFE Settings on the CIFAR10 Dataset (FID score).

These results suggest that our method further improves the performance of existing samplers. Analyses from Figure 1 (left) and Table 2 demonstrate that our method achieves comparable effects to DDIM and DDPM at 50 and 100 NFE settings, respectively, at 25 and 50 NFE settings. This evidences a doubled sampling speed compared to baseline samplers, further validating the effectiveness and orthogonality of our proposed accelerated sampler.

To further verify the efficacy of our method in high-resolution image accelerated sampling, we examined Figure 1 (right). Using high-resolution datasets like LSUN and CelebA-HQ, and setting NFE to 25, we compared the DDPM and our sampler integrated with the momentum mechanism (momentum set at 0.15). Notably, our method produced images with higher fidelity and more intricate texture details in fewer sampling steps. This further substantiates that our method provides efficient accelerated sampling when processing high-resolution datasets.

Ablation Experiments

Momentum-based Accelerated Sampler. In the Methods section, we introduced three momentum mechanisms: momentum, Nesterov momentum, and Adam. To determine the optimal momentum-based accelerated sampling strategy, we conducted a search for the best momentum strategy on the DDPM benchmark, setting the NFE to 100. Notably, in the Adam method, the accumulation of v_t is consistent with the momentum accumulation fraction. Therefore, we directly defined the hyperparameter μ_1 as equal to the optimal parameter m found under the same benchmark method and NFE setting. As shown in Figure 2, the results indicate that the Adam strategy performed the worst, while the pure momentum strategy demonstrated superior performance compared to the Nesterov momentum strategy. Consequently, in other experimental settings, we chose to use only the momentum mechanism to reuse scores for accelerated sampling.

Hyper-Parameters. To determine the optimal range of the hyperparameter m in the momentum method, we searched for the best range of m for all NFE settings in the DDPM method. As shown in Figure 3 (top), we observed that the sample quality initially increases and then decreases with growing momentum. Additionally, we found that the optimal value of m is positively correlated with NFE. Further analysis revealed that a larger m value means reusing more scores. Moreover, a higher NFE represents a baseline sampling with higher fidelity and finer texture details. To further enhance the sampling quality on an already high-quality sampling base, it is necessary to reuse more scores. Therefore, a higher NFE requires a larger hyperparameter m. As seen in Figure 3 (bottom), the sample quality first increases and then decreases with the increasing value of m, indicating that there is a balanced range for reusing scores in different methods and NFE settings to achieve the best effect in detail texture and overall optimization. This finding also suggests that the reused scores (momentum) contain more detailed texture information.

Conclusion

In this paper, inspired by SGD and Langevin dynamics sampling, we proposed a momentum-based score reutilization strategy for accelerated sampling, which does not require any additional training process. Through extensive experimental validation, we found that the method proposed in this paper further enhances the sampling speed on the basis of existing accelerated sampling algorithms, effectively alleviating the current limitations in sampling efficiency of DPMs. This aspect has garnered widespread attention in both academic and industrial circles. In our future work, we plan to apply and extend the proposed method on more advanced accelerated sampling strategies, with the aim of further improving the sampling efficiency of DPMs and driving the development of this field.

References

Bao, F.; Li, C.; Zhu, J.; and Zhang, B. 2022. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffu-

sion probabilistic models. *arXiv preprint arXiv:2201.06503*.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1): 53–65.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Li, L.; Li, H.; Zheng, X.; Wu, J.; Xiao, X.; Wang, R.; Zheng, M.; Pan, X.; Chao, F.; and Ji, R. 2023. AutoDiffusion: Training-Free Optimization of Time Steps and Architectures for Automated Diffusion Model Acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7105–7114.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018): 11.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022a. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022b. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

Nesterov, Y. E. 1983. A method for solving the convex programming problem with convergence rate $Obigl(fracl1k^2bigr)$. In *Dokl. akad. nauk Sssr*, volume 269, 543–547.

Polyak, B. T. 1964. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5): 1–17.

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

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

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency models. *arXiv preprint arXiv:2303.01469*.

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.; Dinh, A.-D.; Liu, D.; and Xu, C. 2023. Boosting diffusion models with an adaptive momentum sampler. *arXiv preprint arXiv:2308.11941*.

Welling, M.; and Teh, Y. W. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 681–688.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zheng, H.; Nie, W.; Vahdat, A.; Azizzadenesheli, K.; and Anandkumar, A. 2023a. Fast sampling of diffusion models via operator learning. In *International Conference on Machine Learning*, 42390–42402. PMLR.

Zheng, K.; Lu, C.; Chen, J.; and Zhu, J. 2023b. DPM-Solver-v3: Improved Diffusion ODE Solver with Empirical Model Statistics. *arXiv preprint arXiv:2310.13268*.

Zhou, Z.; Chen, D.; Wang, C.; and Chen, C. 2023. Fast ODE-based Sampling for Diffusion Models in Around 5 Steps. *arXiv preprint arXiv:2312.00094*.