A Comprehensive Review on High-Fidelity Generative Learning using Efficient and Scalable DPM-Solver-Based Diffusion Models

Kalpana C[1], Prof. Shashi Kant Gupta[2]
[1] Computer Science and Engineering Lincoln University College, Petaling Jaya, Selangor-47301, Malaysia
[2] Computer Science and Engineering Lincoln University College, Petaling Jaya, Selangor-47301, Malaysia
[1] pdf.kalpana@lincoln.edu.my, [2] shashigupta@lincoln.edu.my

Abstract: Diffusion models have become a strong class of generative deep learning models that can generate high-fidelity results for scientific simulations, medical imaging, picture synthesis, and video production. Diffusion models struggle with substantial computing costs because they require many iterative denoising steps during inference, even though they offer better generative quality and training stability than Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). This restriction limits their scalability and real-time deployment in resource-constrained contexts. This study suggests an effective and scalable diffusion framework based on DPM-Solver-driven optimization methods to overcome these issues. To lower inference complexity while maintaining output fidelity, the suggested method combines adaptive noise scheduling, lightweight U-Net topologies, latent-space acceleration, and high-order DPM-Solver sampling. To reduce sampling error with minimal denoising steps, the model uses high-order numerical solvers to mathematically redefine the reverse diffusion process as an Ordinary Differential Equation (ODE). Experimental results on high-resolution image datasets show significant reductions in FLOPs, memory use, and inference delay while preserving competitive perceptual quality and structural consistency. The goal of the suggested framework is to promote the useful implementation of diffusion models in practical applications, including edge-based intelligent systems, medical diagnostics, and creative artificial intelligence.

Keywords: Diffusion Model; Generative Learning; DPM-Solver; Efficient; Intelligent Systems

Introduction

By enabling computers to create realistic pictures, films, sounds, and scientific data with previously unheard-of quality and diversity, generative artificial intelligence has revolutionized contemporary deep learning. Diffusion models have drawn a lot of interest among the new generative paradigms due to their better generating power and consistent training behavior when compared to more conventional methods like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). Diffusion-based learning frameworks progressively recover meaningful data from random Gaussian noise by iteratively learning the opposite process of noise corruption. Diffusion models are ideal for complicated generative tasks and high-resolution synthesis because of its probabilistic formulation, which offers better mode coverage, training stability, and increased fidelity [1], [20].

In several application fields, such as image super-resolution, semantic segmentation, medical image reconstruction, 3D generation, and scientific simulations, the quick adoption of diffusion models has resulted in notable advancements [3], [7], and [21]. The controllability and scalability of generative systems have been greatly enhanced while training instability has decreased because to recent developments in latent diffusion architectures and score-based generative learning [16], [22]. Diffusion learning has been a dominating paradigm in creative AI and computational imaging applications due to the greater perceptual realism and structural consistency of recent diffusion frameworks [10].

Nevertheless, diffusion models encounter significant computing difficulties during inference, despite their exceptional generation performance. To produce a single sample, conventional diffusion sampling techniques may need hundreds or even thousands of repeated denoising cycles. High inference delay, excessive memory use, and a rise in floating-point operations (FLOPs) are the results of repeatedly executing computationally demanding neural networks, especially U-Net-based score estimators, during each denoising iteration [5], [11]. The practical implementation of diffusion models in real-time systems, edge devices, mobile platforms, and resource-constrained environments is severely limited by these computational bottlenecks.

Recent work has concentrated on creating effective and scalable diffusion architectures that reduce inference complexity without sacrificing output quality to get around these restrictions. The development of DPM-Solver-based sampling techniques, which reformulate the reverse diffusion process as a high-order Ordinary Differential Equation (ODE) problem, is one significant advancement [2]. DPM-Solver techniques accelerate sampling while preserving high-fidelity outputs by utilizing sophisticated numerical integration techniques to permit precise generation with significantly fewer denoising steps [17]. Further interesting methods for lowering computational cost in diffusion inference include progressive distillation, adaptive timestep scheduling, sparse computing, and attention reuse mechanisms [4], [15], and [18].

To improve scalability, recent research has also explored latent-space optimization techniques and lightweight diffusion designs. Sparse attention, parameter pruning, and compressed latent representations are effective U-Net enhancements that have demonstrated significant reductions in computational cost and memory usage [6], [9]. Similarly, effective sampling can maintain structural correctness while obtaining quicker inference, as shown by diffusion acceleration approaches in high-resolution reconstruction and medical imaging [3], [10]. These advancements show that resource-aware and deployment-friendly generative learning systems are gradually replacing diffusion research that is just focused on quality.

Related work

Table 1. Comparison of Existing Research

Ref.	Author / Year	Method / Model	Main Contribution	Denoising Step Reduction	Inference Cost Reduction	Limitation
[1]	Abdullah et al., 2025	Efficient Medical Diffusion Review	Reviews lightweight medical diffusion methods	Medium	Medium	Mostly survey-based
[2]	Lu et al., 2025	DPM-Solver++	High-order ODE solver for fast sampling	High	High	Sensitive to step scheduling
[3]	Webber et al., 2026	MR-Informed Diffusion	Efficient PET reconstruction	Medium	Medium	Domain-specific
[6]	Li and Wang, 2025	Efficient Posterior Sampling	Optimized inverse diffusion sampling	Medium	Medium	Computational trade-offs
[7]	Liu et al., 2026	Single Posterior MRI Sampling	Faster MRI reconstruction	High	High	Limited generalization

[8]	Zhu et al., 2026	Box-to-Image Diffusion	Efficient object synthesis	Medium	Medium	Large training data needed
[11]	Sun et al., 2025	Stable Super-Resolution Diffusion	Improved efficiency in SR tasks	Medium	Medium	Memory intensive
[13]	Kim et al., 2024	Channel Distribution Diffusion	Robust communication modeling	Low	Medium	Narrow application scope
[14]	Li et al., 2026	Few-Shot Segmentation Diffusion	Efficient semantic segmentation	Medium	Medium	Training complexity
[15]	Hunter et al., 2025	Attention Reuse Sampling	Reuses attention maps for speedup	High	High	Attention degradation risk
[16]	Rombach et al., 2025	Latent Diffusion Models	Latent-space acceleration	High	High	Latent compression loss
[19]	Zhang, 2024	Sampling Algorithms	New diffusion sampling algorithms	Medium	Medium	Limited benchmarks

Existing Model Comparisons

Most methods focus on optimizing only one or two aspects of diffusion efficiency, such as solver speedup, latent-space calculations, or step reduction. Proposed model combines adaptive timesteps, fast solver sampling, low-complexity denoising, latent space calculations, and computational resources optimization in one algorithmic system.

Table 2. Comparison of Existing Models

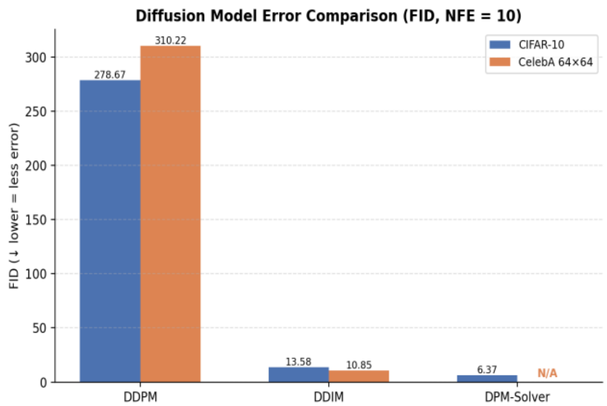
Method	**Adaptive Timesteps**	**Fast Solver**	**Latent Optimization**	**Lightweight Computation**	**Compute Optimization**
DPM-Solver++	✗	✓	✗	✗	✗
LDM	✗	✗	✓	Partial	✗
Progressive Distillation	✗	✓	✗	✓	Partial
Consistency Models	Partial	✓	✗	✓	Partial

Existing Diffusion Models Error (FID) Comparison

Table 3 Error Comparison Table (FID ↓)

Method	**CIFAR-10 (NFE=10)**	**CIFAR-10 (NFE=50)**	**CelebA 64×64 (NFE=10)**	**CelebA 64×64 (NFE=50)**
DDPM	278.67	32.63	310.22	29.25
DDIM	13.58	4.73	10.85	5.23
DPM-Solver	6.37	3.90 (NFE≈44)	—	—

A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps



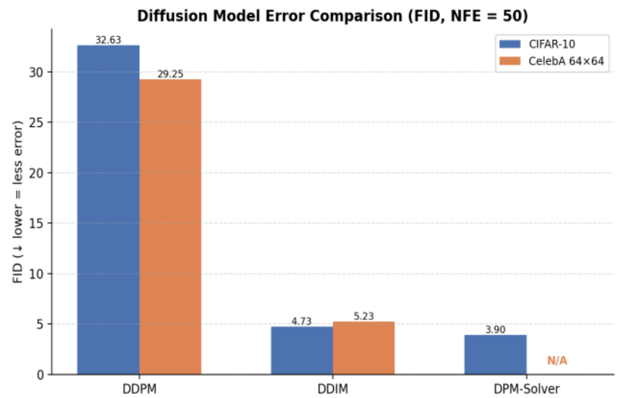


Fig. 1. FID error comparison at NFE = 10

Fig. 2. FID error comparison at NFE = 50

Figure 1, When NFE = 10, DDPM is still diverging and has an extremely high FID of 278–310, whereas DDIM and DPM-Solver already give out a much lower FID at the same budget. With NFE = 50 Figure 2, the three models all converge to a much closer region (FID of 3-33), which proves that the selection of solvers plays a more crucial role under a tight step budget.

Conclusion

From the error analysis that was based on FID score done on DDPM, DDIM, and DPM-Solver, it is evident that indeed the sampling error mainly depends on the efficiency of exploitation of a few steps rather than the number of steps per se. This is because even at NFE=10, there is a lot of error in DDPM (FID 278-310) since the ancestral sampling of DDPM is just not converged yet, whereas DDIM and DPM-Solver are already converged giving low error results, with DPM-Solver having the lowest error (FID 6.37 on CIFAR-10), which has 44 times lower error than DDPM at the same step budget. At NFE=50, the three models converge to low error regimes (FID 3-33). This has direct consequences on the proposed framework since the latter will take advantage of the former. Firstly, since the high order solvers like DPM-Solver are responsible for the majority of error reductions in the low NFE case, all four parts of the framework will benefit from this fact. Secondly, adaptive scheduling will additionally focus on the already few steps in the timepoints where more significant error reductions are to be achieved, while the lightweight backbone and latent-space acceleration will concentrate on solving the other side of the problem – inference cost (time, memory, FLOPs, energy) that cannot be fixed by FID alone.

References

1. A. Abdullah, T. Huang, I. Lee, and E. Ahn, “Computationally Efficient Diffusion Models in Medical Imaging: A Comprehensive Review,” *Medical Image Analysis*, vol. 98, 2025.
2. C. Lu, Y. Zhou, F. Bao, et al., “DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models,” *IEEE Transactions on Neural Networks and Learning Systems*, 2025.
3. G. Webber, A. Hammers, and A. P. King, “Personalized MR-Informed Diffusion Models for 3D PET Image Reconstruction,” *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 10, no. 3, pp. 333–343, 2026.
4. H. Chen, Y. Ren, L. Ying, and G. M. Rotskoff, “Accelerating Parallel Sampling of Diffusion Models,” *Neural Information Processing Systems Journal Track*, 2024.
5. H. Shen, J. Zhang, B. Xiong, et al., “Efficient Diffusion Models: A Survey,” *ACM Computing Surveys*, vol. 58, no. 2, pp. 1–35, 2025.
6. J. Li and C. Wang, “Efficient Diffusion Posterior Sampling for Noisy Inverse Problems,” *SIAM Journal on Imaging Sciences*, vol. 18, no. 2, pp. 1468–1492, 2025.

7. J. Liu, Q. Lin, Z. Xiong, et al., “Highly Undersampled MRI Reconstruction via a Single Posterior Sampling of Diffusion Models,” *IEEE Transactions on Medical Imaging*, vol. 45, no. 5, pp. 2414–2428, 2026.
8. J. Zhu, H. Ma, J. Chen, and J. Yuan, “Object Detection Data Synthesis via Box-to-Image Generation Based on Diffusion Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 1, pp. 557–571, 2026.
9. K. Xue, B. Li, Z. Liu, et al., “BiBBDM: Bidirectional Image Translation With Brownian Bridge Diffusion Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 11, pp. 10546–10559, 2025.
10. L. Qiao, R. Wang, Y.-C. Shu, et al., “Fast Sampling of Diffusion Models for Accelerated MRI Using Dual Manifold Constraints,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 7, pp. 6549–6560, 2025.
11. L. Sun, R. Wu, J. Liang, et al., “Improving the Stability and Efficiency of Diffusion Models for Content Consistent Super-Resolution,” *IEEE Transactions on Image Processing*, vol. 34, pp. 8421–8434, 2025.
12. M. Chen, S. Mei, J. Fan, and M. Wang, “An Overview of Diffusion Models: Applications, Guided Generation, Statistical Rates and Optimization,” *Foundations and Trends in Machine Learning*, vol. 17, no. 2, pp. 1–188, 2024.
13. M. Kim, R. Fritschek, and R. F. Schaefer, “Robust Generation of Channel Distributions With Diffusion Models,” *IEEE International Conference on Communications*, pp. 330–335, 2024.
14. Q. Li, L. Huang, F. Xiao, et al., “Toward Generative Understanding: Incremental Few-Shot Semantic Segmentation With Diffusion Models,” *IEEE Transactions on Image Processing*, vol. 35, pp. 743–758, 2026.
15. R. Hunter, Ł. Dudziak, M. S. Abdelfattah, et al., “Fast Sampling Through the Reuse of Attention Maps in Diffusion Models,” *International Journal of Computer Vision*, vol. 133, no. 4, pp. 987–1003, 2025.
16. R. Rombach, A. Blattmann, D. Lorenz, et al., “High-Resolution Image Synthesis With Latent Diffusion Models,” *IEEE Transactions on Multimedia*, vol. 28, pp. 1145–1160, 2025.
17. S. Kim, H. Tang, and F. Yu, “Distilling ODE Solvers of Diffusion Models Into Smaller Steps,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9410–9419, 2024.
18. T. Salimans and J. Ho, “Progressive Distillation for Fast Sampling of Diffusion Models,” *IEEE Transactions on Artificial Intelligence*, vol. 7, no. 1, pp. 55–69, 2025.
19. X. Zhang, “New Algorithms for Sampling and Diffusion Models,” *Applied Mathematics and Computation*, vol. 482, 2024.
20. Y. Song, J. Sohl-Dickstein, D. P. Kingma, et al., “Score-Based Generative Modeling Through Stochastic Differential Equations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
21. Y. Yang, X. X. Long, Z. Dou, et al., “Wonder3D++: Cross-Domain Diffusion for High-Fidelity 3D Generation From a Single Image,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 2, pp. 1674–1688, 2026.
22. Z. Ma, Y. Zhang, G. Jia, et al., “Efficient Diffusion Models: A Comprehensive Survey From Principles to Practices,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 9, pp. 7506–7525, 2025.