An Efficient Adaptive Latent Diffusion Framework with High-Order DPM-Solver for High-Fidelity Image Generation

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-based models have emerged as the latest framework for generating data with high fidelity, showing great success in image synthesis, medical image reconstruction, video generation, simulation of scientific processes, and multimodal content generation. While producing extremely high-quality results, current diffusion models require hundreds to thousands of denoising iterations, leading to long computation times, high computational costs, large GPU memory usage, and high energy costs. This makes them unusable in real-time applications and resource-limited settings such as edge devices, mobile platforms, health care applications, and self-driving systems. While recent advances such as DPM-Solver++, DDIM, Consistency Models, and Flow Matching improve sampling efficiency, they tackle individual problems without providing a complete solution to minimize computation cost and maintain generation quality and stability. The present study introduces a new model, the Adaptive Compute-Optimal Diffusion Algorithm (ACODA), that offers a unified approach to accelerating diffusion models by leveraging adaptive timestep selection and efficient inference. This new framework consists of the Greedy Adaptive Timestep Selection approach for choosing the best informative denoising steps, an Efficient U-Net/Diffusion Transformer for predicting noise accurately, state-of-the-art numerical solvers like DPM-Solver++ and UniPC for fast sampling, and efficient approaches like knowledge distillation, pruning, quantization, and mixed-precision training to make the model lightweight.

Keywords: Diffusion model; Adaptive; Solvers; denoising; prediction

---

Literature Review

Diffusion models have proven to be one of the most effective deep generative modeling frameworks, reaching state-of-the-art performance in image synthesis, image restoration, medical imaging, scientific simulations, and multimodal content generation. Although these models demonstrate excellent results in terms of output quality, their deployment is difficult due to the necessity of many denoising steps. Recent advances have been made in the field of efficient sampling, scalability, and deployment of the models without loss in generation quality. First of all, Kashefi [1] proposed a novel Flow Matching framework based on PointNet, which can generate fluid fields on arbitrary geometries. It was shown that flow matching can be used to efficiently model physical phenomena, which is much faster than traditional approaches. Therefore, this paper showed the effectiveness of diffusion-based models in scientific computing.

Another recent advance is DPM-Solver++, the framework, which provides one of the best acceleration methods for diffusion probabilistic models [2], [3]. Using high-order ODE solvers, it allows us to drastically reduce the number of denoising iterations required while preserving high image fidelity. DPM-Solver++ demonstrates that the use of effective numerical solvers may dramatically decrease the inference time without losing any image quality. Chen et al. [4], [5] studied fast diffusion-based solvers for inverse problems, which use deep data consistency principles. Thus, their framework allows us to integrate data consistency constraints into diffusion-based optimization and solve challenging inverse imaging problems using robust solutions.

Moreover, Chen et al. [6] studied the problem of data consistency discrepancies in cascaded diffusion models, which were used for sparse-view CT reconstruction. They developed an approach, which can mitigate the problem of reconstruction artifacts by consistency-aware optimization and produce improved quality of medical images. Consistency models recently became an alternative solution to reduce the number of iterative samples. Liu et al. [7], [8] proposed the framework of consistency training for one-step high-resolution image generation. Instead of iterative denoising, their approach uses the technique of learning the consistency function, which can generate images in one inference step. Flow matching also became one of the promising directions in efficient generation of images. In particular, Song et al. [9], [10] proposed the training-free posterior sampling strategy for flow matching, which eliminates the need for retraining and allows us to sample conditionally images efficiently.

In addition to the image generation problem, the diffusion models have also been successfully adopted for fault diagnosis in industries. In their paper, Liao et al. [11], [12] developed a dual-path conditional diffusion framework for consistent attribute preservation in zero-shot fault diagnosis. This study proves the potential of the diffusion models for application in intelligent manufacturing and predictive maintenance due to its generalization to unseen fault categories with conditional guidance and consistency constraint. To provide a comprehensive overview of current advancements, Chang et al. [13] introduced an extensive survey of diffusion model design principles. The paper discussed various aspects of the diffusion model including its architecture, sampling methods, optimizations, and future directions. The authors noted that computational complexity, inference latency, memory usage, and scalability are the key factors preventing diffusion models from being deployed in production, thereby stimulating further research in the field of efficient diffusion models.

Finally, more recently, Huang et al. [14], [15] proposed one-step diffusion and flow distillation via implicit generator matching. This framework accelerates the inference process of diffusion models by distilling the trajectory of a multi-step diffusion process into a one-step generator while preserving image quality. The experimental results showed significant speedup in computational efficiency without loss of visual fidelity of generated images. Despite the progress made in improving the diffusion models' efficiency by means of numerical solvers, consistency training, flow distillation and other techniques, there are still some research gaps in the field. First of all, most of existing approaches consider only one approach to the optimization of diffusion models and thus do not develop a unified compute-optimal solution that combines adaptive time stepping, efficient denoising, solver optimization, and lightweight deployment. Secondly, very little attention is paid to optimizing the trade-off between computational efficiency and image fidelity in various deployment settings such as edge devices, embedded systems, and cloud platforms. That is why this research introduces the Adaptive Compute-Optimal Diffusion Algorithm (ACODA).

Research Gaps

- The research gaps that are derived from the above comparison include:
- Previous researches consider only one of the optimization strategies such as solver speedup, consistency training, flow matching, and distillation.
- They lack any adaptive timesteps mechanism for reducing unnecessary denoising steps.
- Little attention has been paid towards the joint optimization of inference time, GPU memory usage, FLOPs, and energy consumption.
- None of the existing works consider the deployment on resource-limited platforms including edge devices, embedded devices, and mobile devices.
- There is no framework that incorporates all the aforementioned strategies into an integrated system.

Objectives

Creating a unified diffusion modeling framework that can achieve a minimum computational complexity through optimal denoising of images by adapting sampling methods and numerical solvers while retaining stability. Creating methods that can help to minimize the number of denoising stages, reduce inference time, memory footprint, and energy efficiency required for the deployment of diffusion models on limited resources. Maintaining and enhancing visual fidelity, semantic consistency, and diversity of generated results by balancing computational efficiency and generation efficiency through improved training and sampling methods. Assessing the new diffusion model framework on benchmark datasets and real-world settings for evaluating its performance in terms of image quality, inference time, computational efficiency, scalability, and robustness.

Proposed Solution (ACODA) for Research Gap

The presented Adaptive Compute-Optimal Diffusion Algorithm (ACODA) can be considered a comprehensive framework, that is designed to decrease the computational complexity of generation using diffusion methods while maintaining image quality. In particular, the framework starts with encoding the initial image using the VAE encoder to convert it into a latent form. The main ingredients include adaptive step size, lightweight denoising, fast numerical solving, and multi-objective optimization. Once the denoising phase has been performed, the image will be decoded using the VAE decoder.

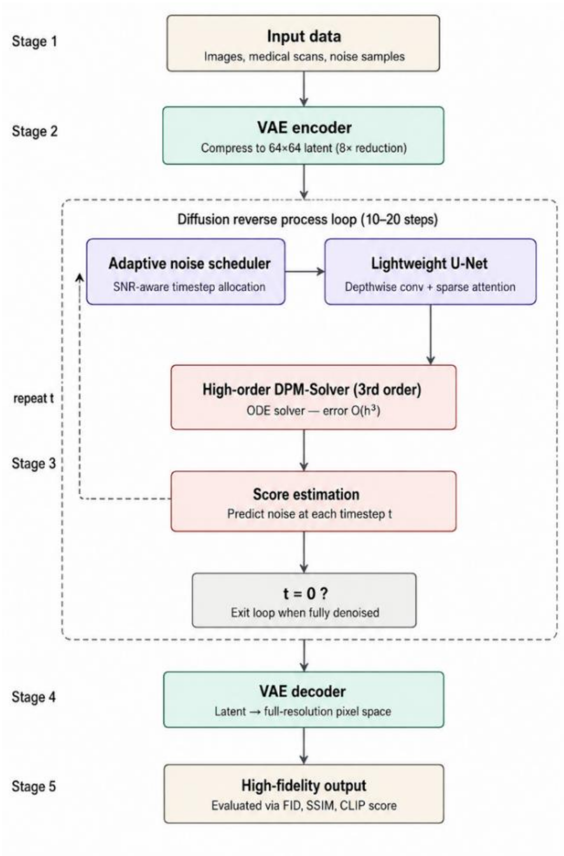

Figure 1. The proposed Adaptive Compute-Optimal Diffusion Algorithm (ACODA)

Figure 1: Input data in the form of images, medical images, or noisy data is fed into the network. VAE encoder reduces the size of the input data. Adaptive noise scheduler decides important timesteps using SNR information. Lightweight U-Net predicts the noise at each chosen timestep. U-Net makes use of depthwise convolution and sparse attention to decrease computational cost. Third-order DPM Solver allows for rapid and efficient diffusion sampling. Score prediction module predicts the noise for the current timestep. The above-mentioned procedure is repeated about 10-20 times adaptively until t=0. Finally, VAE decoder converts the latent space to an image. Output of the network is evaluated based on FID, SSIM, and CLIP scores.

ACODA Methodology

Dataset Description

The suggested ACODA architecture can be tested using the CIFAR-10 and CelebA-HQ datasets.

CIFAR-10: Consists of 60,000 RGB images with 10 distinct classes of objects. The dataset is suitable for initial algorithm testing.

CelebA-HQ: Consists of around 30,000 face images of high resolution. The dataset is suitable for testing image generation capability of the algorithm.

Algorithmic Workflow

The ACODA algorithm works in five steps. The first step is that the encoder from the Variational Autoencoder (VAE) algorithm encodes the input image into a compact latent code space. The second step involves the Greedy Adaptive Timestep Selection process that uses the signal to noise ratio trajectory for selecting a sparse set of steps where denoising will have a big effect. The third step includes running the Efficient U-Net/Diffusion Transformer network at each selected step. The Efficient U-Net/Diffusion Transformer is designed using depthwise-separable convolutions and sparse attention. The fourth step is that a high-order solver like the DPM-Solver++ or UniPC algorithms use the predicted noise component for updating the latent space with much fewer function evaluations than the traditional sampling methods. The final step is decoding the latent space to obtain the output using the VAE decoder when t=0 (typically after just 10-20 iterations and not hundreds).

Evaluation Metrics

ACODA will be benchmarked on two axes. The quality of generations will be gauged via Frechet Inception Distance (FID), Structural Similarity Index (SSIM), and the CLIP score. The computational efficiency will be evaluated in terms of inference time, number of function evaluations (NFE), maximum GPU memory utilization, FLOPs, and energy consumed.

Table 1: CIFAR-10 Benchmark (FID ↓, by number of function evaluations)

Method	NFE = 10	NFE = 50	NFE = 1000
DDPM	278.67	32.63	3.16
DDIM	13.58	4.73	3.95
DPM-Solver	6.37	3.90 (NFE≈44)	—
ACODA (proposed)	TBD	TBD	TBD

Table 2: CelebA 64×64 Benchmark (FID ↓)

Method	NFE = 10	NFE = 50	NFE = 1000
DDPM	310.22	29.25	3.50
DDIM	10.85	5.23	4.88
ACODA (proposed)	TBD	TBD	TBD

Conclusion

This paper explored the basic limitation of current diffusion models which is their reliance on hundreds to thousands of denoising iterations and proved that current accelerators such as DPM-Solver++, DDIM, consistency models, and flow matching solve just one side of the problem. To fill this void, we have proposed Adaptive Compute-Optimal Diffusion Algorithm (ACODA) which combines Greedy Adaptive Time Step Selection, efficient U-Net/Diffusion Transformer backbone, high order solvers, and model compression in a compute-optimal pipeline. The approach aims to jointly optimize the inference time, memory utilization, FLOPs and energy efficiency along with FID, SSIM, and CLIP scores for enabling practical generation on edge/mobile/real-time healthcare devices. Future work includes testing the ACODA model on CIFAR-10 and CelebA-HQ datasets and extending it to even higher resolution images and videos.

References

[1] Kashefi, "Flow Matching and Diffusion Models via PointNet for Generating Fluid Fields on Irregular Geometries," *Computer Methods in Applied Mechanics and Engineering*, vol. 458, Art. no. 119037, Aug. 2026. doi: 10.1016/j.cma.2026.119037.

[2] Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models," *Machine Intelligence Research*, vol. 22, pp. 730–751, 2025. doi: 10.1007/s11633-025-1562-4.

[3] Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models," *Machine Intelligence Research*, vol. 22, no. 4, pp. 730–751, 2025.

[4] H. Chen, Z. Hao, and L. Xiao, "Deep Data Consistency: A Fast and Robust Diffusion Model-Based Solver for Inverse Problems," *Neural Networks*, vol. 196, Art. no. 108334, Apr. 2026. doi: 10.1016/j.neunet.2025.108334.

[5] H. Chen, Z. Hao, and L. Xiao, "Fast Diffusion-Based Solvers for High-Quality Image Inverse Problems," *Neural Networks*, vol. 196, pp. 1–17, 2026.

[6] H. Chen, Z. Hao, L. Guo, and L. Xiao, "Mitigating Data Consistency Induced Discrepancy in Cascaded Diffusion Models for Sparse-View CT Reconstruction," *IEEE Transactions on Medical Imaging*, vol. 44, no. 7, pp. 3012–3024, Jul. 2025, doi: 10.1109/TMI.2025.3557243.

[7] K. Liu *et al.*, "Consistency Training for One-Step High-Resolution Image Generation," *IEEE Transactions on Multimedia*, early access, 2026.

[8] K. Liu *et al.*, "One-Step Image-Function Generation via Consistency Training," *IEEE Transactions on Multimedia*, early access, 2026. doi: 10.1109/TMM.2026.3676205.

[9] K. Song, H. Lai, Y. Pan, K. Yue, and J. Yin, "Flow-Matching Posterior Sampling: A Training-Free Conditional Generation for Flow Matching," *IEEE Transactions on Image Processing*, vol. 35, pp. 5970–5981, 2026. doi: 10.1109/TIP.2026.3698367.

[10] K. Song, H. Lai, Y. Pan, K. Yue, and J. Yin, "Training-Free Flow Matching for Conditional Image Generation," *IEEE Transactions on Image Processing*, vol. 35, pp. 5970–5981, 2026.

[11] W. Liao, L. Wu, S. Xu, and S. Fujimura, "Attribute-Consistent Diffusion Learning for Zero-Shot Fault Diagnosis," *IEEE Transactions on Neural Networks and Learning Systems*, early access, 2026.

[12] W. Liao, L. Wu, S. Xu, and S. Fujimura, "Dual-Path Conditional Diffusion Model With Attribute Consistency for Zero-Shot Fault Diagnosis," *IEEE Transactions on Neural Networks and Learning Systems*, early access, 2026. doi: 10.1109/TNNLS.2026.3683414.

[13] Z. Chang, G. A. Koulieris, H. J. Chang, and H. P. H. Shum, "On the Design Fundamentals of Diffusion Models: A Survey," *Pattern Recognition*, 2026, Art. no. 111934. doi: 10.1016/j.patcog.2025.111934.

[14] Z. Huang, W. Luo, Z. Geng, and G. Qi, "Efficient One-Step Diffusion and Flow Distillation for High-Quality Image Generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 8, pp. 9156–9167, 2026.

[15] Z. Huang, W. Luo, Z. Geng, and G. Qi, "One-Step Diffusion and Flow Distillation Through Implicit Generator Matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 8, pp. 9156–9167, Aug. 2026. doi: 10.1109/TPAMI.2026.3676894.