Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37251

A Systematic Comparison and Fusion of Generative Adversarial Networks and Diffusion Models for Image Generation

Zheping Ding

Lassonde School of Engineering, York University, Toronto, Canada
zheping2@my.yorku.ca

Abstract. Generative Adversarial Networks (GANs) and diffusion models are two mainstream approaches in image generation. Although their underlying principles differ, both can be unified under the framework of particle models (PMs). GANs enable fast generation but suffer from unstable training and mode collapse, whereas diffusion models produce high-quality, diverse samples at the cost of slow sampling. Although various diffusion model variants have been proposed to accelerate sampling, a systematic comparison under identical experimental conditions remains lacking. This paper presents a fair, comprehensive comparison of GAN, Denoising Diffusion Probabilistic Model (DDPM), Denoising Diffusion Implicit Model (DDIM), and a proposed fusion model on the Fashion-MNIST dataset, evaluating generation quality with Fréchet Inception Distance (Feature-FID) and Kernel Inception Distance (Feature-KID) and analyzing the effects of sampling steps and stochasticity. Results show that DDIM with stochastic sampling (η=1) achieves the lowest KID of 575.51±5.69, outperforming deterministic DDIM and the standard GAN. The proposed fusion model, combining GAN warm-up with DDIM refinement, reduces the number of sampling steps while achieving a KID of 672.17±65.15, offering a trade-off between generation quality and inference efficiency.

Keywords: GAN, diffusion models, fusion model, hybrid model, Fashion-MNIST

1. Introduction

Both diffusion models and GANs, introduced by Goodfellow et al. in 2020, can be used for image generation [1]. However, the research result of Zhu, by comparing GANs and score-based generative models on the CIFAR-10 dataset, illustrated that score-based models outperformed GANs in terms of sample quality and diversity [2]. Although the traditional score-based generative model DDPM achieves better image quality and diversity, it requires more computational power [2]. In contrast, GANs offer faster generation speed.

Researchers have proposed various improved variants of DDPM. To enhance sampling speed, Song et al. (2020) introduced DDIM, which achieves faster generation through deterministic sampling with η=0 [3]. Lam et al. (2021) proposed Bilateral Denoising Diffusion Models (BDDM),

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37251

which introduce a scheduling network φ to learn short timestep schedules, thereby accelerating the denoising process [4].

Although these models have reported results in their respective papers, they differ in network architectures, training epochs, and hardware configurations, making direct cross-paper comparisons unreliable. There is a lack of systematic horizontal comparison of various diffusion model variants against GANs under identical experimental conditions.

This paper investigates the performance gap between three diffusion-based generative models and GANs in terms of generation quality. According to Franceschi et al. (2023), GANs and diffusion models are not mathematically opposed but rather different instances within a unified particle model framework [5]. This paper therefore attempts to combine GANs and diffusion models to obtain an optimized generative network.

This study systematically compares GAN, DDPM, and DDIM on Fashion-MNIST under identical experimental conditions, showing that stochastic DDIM sampling (η=1) outperforms deterministic sampling (η=0), reaching a KID of 575.51±5.69 with 20 sampling steps. Building on the particle-model perspective, a hybrid strategy combining GAN warm-up with DDIM refinement is further proposed, reducing the number of sampling steps while achieving a KID of 672.17±65.15, though its generation quality remains below that of the best diffusion configuration.

2. Related work

2.1. Diffusion models

DDPM progressively add Gaussian noise to data through a forward Markov chain $q(x_t \mid x_{t-1})$, with the noising process defined as:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} z_t \tag{1}$$

A neural network $\epsilon_\theta(x_t t)$ is trained to predict the added noise during the reverse process, thereby recovering the model distribution $p_\theta(x_0)$ [6]. The training objective is to make $p_\theta(x_0)$ as close as possible to the true data distribution $q(x_0)$, simplified as the mean squared error $\| \epsilon_\theta(x_t t) - \epsilon \|^2$ for noise prediction. DDPM achieves high image generation quality, but its sampling process requires T sequential steps (typically $T = 1000$), making it significantly slower than single-step generative models like GANs. Huang et al. (2024) showed that DDPM can be viewed as an adaptively discretized stochastic differential equation (SDE), and that it automatically adapts to the intrinsic low-dimensional structure of data, achieving optimal sampling efficiency in theory [7]. This suggests that rapidly pulling data back to the low-dimensional manifold can accelerate the model.

To alleviate the slow sampling problem of DDPM, Song et al. (2020) proposed DDIM [3]. DDIM modifies the decomposition rule of the forward noising process. For s time steps, the diffusion chain is decomposed as:

$$q(x_{1:s} \mid x_0) = q(x_s \mid x_0) \prod_{i \in A} q(x_{i-1} \mid x_i,\ x_0) \prod_{j \in B} q(x_j \mid x_0) \tag{2}$$

And the reverse diffusion chain is decomposed as:

DOI: 10.54254/2755-2721/2026.37251

$$P(x_{0:s}) = P(x_s)\prod_{i\in A} P(x_{i-1} \mid x_i)\prod_{j\in B} P(x_0 \mid x_j) \tag{3}$$

This ensures the training objective remains unchanged while enabling skip-step sampling. The loss function satisfies:

$$\mathrm{KL}(q(x_{i-1} \mid x_i,\ x_0) \parallel P(x_{i-1} \mid x_i)) \propto \parallel \epsilon_\theta(x_i i) - \epsilon_i \parallel^2 \tag{4}$$

Which is consistent with the optimization objective of DDPM.

2.2. GAN model

GANs consist of two components: a generator G and a discriminator D. The generator G maps a latent variable $z \sim p_z$ to the image space to produce a generated sample $G(z)$. The discriminator D takes as input both real images $x \sim q(x)$ and generated samples $G(z)$, and outputs a probability value to distinguish between the two. The discriminator D and the generator G are optimized through continuous adversarial training. The ultimate objective of this adversarial game can be formulated as the following minimax objective function:

$$\min_G \max_D \mathbb{E}_{x\sim q(x)}[logD(x)] + \mathbb{E}_{z\sim p_z}[log(1 - D(G(z)))] \tag{5}$$

At the optimal solution of this objective $p_G = q$.

2.3. Advantages and disadvantages

GANs and diffusion models differ fundamentally in their generative mechanisms. Shen et al. (2024) demonstrated on the MNIST dataset that GANs, with their single-step generation mechanism, are significantly faster in inference than diffusion models. However, their generation quality and diversity fall short of those achieved by diffusion models [8].

3. Methodology

3.1. Dataset and preprocessing

The experiments are conducted on the Fashion-MNIST grayscale image dataset (10 classes), following the official split of 60,000 training images and 10,000 test images [9]. To evaluate model robustness under different data scales, this paper additionally performs class-wise stratified sampling to create training subsets at different proportions (e.g., 50%, 70%, 90% of the training data) for comparison. All models adopt the same data preprocessing pipeline: pixel values are normalized from [0, 255] to [-1, 1] to match the tanh activation range of the generator output layer. Batch size and training epochs are also kept identical across all models to ensure fair comparison.

Three representative generative models are selected as baselines: GAN, DDPM, and DDIM (an accelerated variant of DDPM). All models are implemented with a unified U-Net architecture consisting of 4 downsampling/upsampling stages, an initial channel size of 64, channel multipliers

DOI: 10.54254/2755-2721/2026.37251

{1, 2, 4, 8}, a time embedding dimension of 256, and self-attention mechanisms introduced at the 16×16 resolution layers, to control for architectural variables.

3.2. Experimental setup and evaluation protocol

Table 1 summarizes the training hyperparameters for the three generative models, including Batch size, Learning Rate, Optimizer, Epochs, and Device.

Table 1. Training hyperparameters for different generative models

name	GAN	DDPM	DDIM
Batch Size	128	128	128
Learning Rate	2e-4	2e-4	2e-4
Optimizer	Adam	AdamW	AdamW
Epochs	50	49	-

All experiments are conducted with a single fixed set of three random seeds (seeds = 42, 123, 456), and no multi-seed repetition is performed.

3.3. Fusion model design

3.3.1. Preliminary: particle models and existing instantiations

The preliminary framework proposed by Franceschi et al. is reviewed first. They proposed a unified framework for diffusion models and GANs, showing that they are based on similar particle evolution equations. Particle models follow a gradient vector. A generator can be trained using score-based gradients as an alternative to adversarial training (Score GAN). Meanwhile, a GAN can also be trained without a generator, using only the discriminator to synthesize samples.

Score-based diffusion models belong to the class of PMs and leverage Wasserstein gradient flow. The differential equation governing the position of each particle is given by:

$$dx_t = -\nabla_w F(\rho_t)(x_t)dt \tag{6}$$

DDPM, DDIM, and Noise Conditional Score Network(NCSN) are mathematically equivalent, and all belong to the class of PMs, satisfying Equation 6 [10]. Under the unified framework, GANs follow the same particle evolution paradigm as DDPM, but the driving force is provided by an equivalent discriminator.

3.3.2. Design and implementation of the proposed fusion model

During inference, first sample random noise $z \sim \mathcal{N}(0I)$ and feed it into the trained conditional GAN to generate the intermediate state $\widehat{x}_{t_K} = G_\theta(z)$ in a single forward pass. This intermediate state is then used as the initial value for DDIM, which performs deterministic iterative denoising along the probability flow Ordinary Differential Equation (ODE) starting from t_K, ultimately producing the final image x_0. This pipeline compresses the diffusion timesteps from T steps to $(T - t_K)$ steps, thereby achieving acceleration.

DOI: 10.54254/2755-2721/2026.37251

The equivalent discriminator loss of the fusion model fuse cannot be exactly the same as $\mathscr{L}_{\mathrm{GAN}}$, because the goal of the equivalent discriminator is to make the input distribution ρ for DDIM easy to denoise, rather than only caring about the generated samples $G(z)$.

$$d\theta_t = -\nabla_\theta \mathbb{E}_{z\sim p_z}[h\rho GAN(G_\theta(z))]\, dt = -\nabla_\theta \mathbb{E}_{z\sim p_z}[-c \circ f_\rho(G_\theta(z))]\, dt \tag{7}$$

Where f_ρ is the discriminator and c is a function that maps the discriminator output to the loss space.

A novel fusion model that redesigns the driving force is proposed. Instead of using the original equivalent discriminator, introduce a hybrid loss function that combines adversarial training with a reconstruction constraint. Proposed loss is defined as:

$$\mathscr{L}_{fuse} = \mathscr{L}_{GAN} + \lambda \cdot \mathbb{E}_{z\sim p_z}[\parallel DDIM(G_\theta(z)) - x_0 \parallel^2] \tag{8}$$

The corresponding particle evolution (in the generator parameter space) for Equation 8 becomes:

$$d\theta_t = -\nabla_\theta(\mathscr{L}_{\mathrm{GAN}} + \lambda \cdot \mathbb{E}_{z\sim p_z}[\parallel \mathrm{DDIM}(G_\theta(z)) - x_0 \parallel^2])dt \tag{9}$$

Where λ is a hyperparameter. Based on the above definition, the following can be obtained:

$$h_\rho = \begin{cases} -c \circ f_\rho, & t \in [tK, T) \\ log\left(p_{data} \star k^\sigma_{RBF}\right) - log\left(\rho_t\right), & t \in \left[0, tK\right) \end{cases} \tag{10}$$

The original framework uses a single equivalent discriminator $-c \circ f_\rho$ as the driving force.

New proposed fusion model replaces this driving force with a hybrid gradient that combines L_{GAN} and a reconstruction loss, enabling the generator to simultaneously optimize for both realism and reconstructability.

Since the DDIM sampling process consists of dozens of non-linear transformation steps, propagating gradients through the entire sampling chain could cause severe gradient vanishing or explosion issues. To address this, employ a stop-gradient strategy that truncates the computational graph at the output of DDIM. Specifically, detach the DDIM-refined output x_{final} and compute the reconstruction loss as:

$$\mathscr{L}_{\text{recon}} = \parallel G_\theta(z) - \mathrm{sg}(x_{\text{final}}) \parallel_2^2 \tag{11}$$

Where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. This strategy provides the generator with a stable optimization objective by truncating the long-range computational graph. Stable Target Field (STF) effectively reduces the variance of the training objective in diffusion model score matching by introducing a reference batch to compute weighted conditional scores, thereby improving training

DOI: 10.54254/2755-2721/2026.37251

stability and convergence speed [11]. At the gradient regulation level, Adaptive Group-wise Gradient Clipping (AGGC) simultaneously mitigates both gradient explosion and vanishing issues by grouping parameters according to their functional types and adaptively adjusting the clipping intervals [12]. Collectively, these works demonstrate that ensuring optimization stability—whether through constructing low-variance objectives or adaptively regulating gradients—is a crucial pathway to improving the training effectiveness of deep generative models.

This yields a new fusion model design within the PM framework. From an implementation perspective, the fusion model replaces the first few diffusion timesteps of DDIM with a hybrid model (fuse) to pull particles back to the low-dimensional manifold more quickly. The loss function is replaced by $\mathscr{L}_{\text{fuse}}$.

3.4. Evaluation metrics

Quality was measured by Feature-FID and Feature-KID in a 128-dimensional feature space from a project-trained Fashion-MNIST Convolutional Neural Network (CNN). Let (μ_r, Σ_r) and (μ_g, Σ_g) denote empirical feature moments for real and generated samples. The implemented Fréchet feature distance is:

$$\text{FeatureFID} = \| \mu_r - \mu_g \|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \tag{12}$$

Feature-KID uses the cubic polynomial kernel:

$$k(uv) = \left(\frac{u^T v}{d} + 1\right)^3,\ d = 128 \tag{13}$$

And an unbiased squared maximum mean discrepancy estimator:

$$\widehat{\text{MMD}}_u^2 = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k(r_i r_j) + \frac{1}{m(m-1)} \sum_{i \neq j}^{m} k(g_i g_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(r_i g_j) \tag{14}$$

The reported Feature-KID is the mean over repeated random subsets. These metrics support controlled comparisons within this project only; their numerical scales are not interchangeable with standard ImageNet Inception-space FID or KID.

3.5. Experiments

Table 2 divides the experiments into six groups, each representing a specific model configuration that covers different generative models and sampling strategies.

Table 2. Experimental groups and their diagnostic purposes

Group	Model and sampler	NFE	η
B100	T=100 ancestral	100	-
B20	Independently trained T=20 ancestral	20	-

DOI: 10.54254/2755-2721/2026.37251

Table 2. (continued)

D100	T=100 DDIM, adjacent steps	100	0
D20	T=20 DDIM, adjacent steps	20	0
D0	T=100 DDIM, adjacent steps	100	1
D1	T=20 DDIM, adjacent steps	20	1

4. Results and analysis

4.1. Results

Table 3 reports the FID and KID of different experimental configurations, all evaluated on the same test set. With B100 as the baseline, D1 achieves the lowest FID and KID of 478.38 ± 5.60 and 575.51 ± 5.69, respectively, showing a slight improvement over the baseline. Among the diffusion models, D20 performs the worst. GAN exhibits the poorest generation quality among all models.

Table 3. FID and KID results across sampling configurations (mean ± std over seeds 42, 123, 456)

name	FID	KID
B100	520.38 ± 22.10	563.66 ± 26.00
B20	486.06 ± 2.75	565.66 ± 5.03
D100	591.22 ± 17.13	618.46 ±7.86
D20	593.25 ± 15.78	620.55 ±9.81
D0	486.96 ± 4.04	576.04 ± 3.71
D1	478.38 ± 5.60	575.51 ± 5.69
GAN	1002.74	829.33

Figure 1 presents the FID of models with different sampling steps, demonstrating that a sampling step of 20 is the optimal configuration for DDIM in these experiments, achieving the lowest FID.

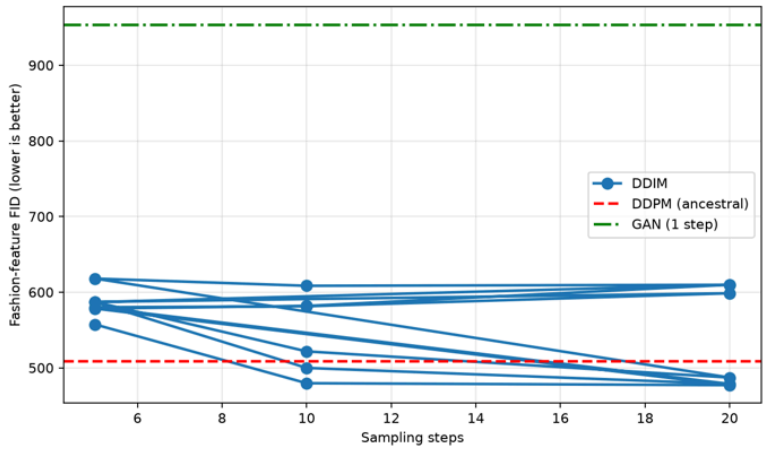

Figure 1. DDPM vs DDIM vs GAN: quality vs sampling steps (photo credit: original)

DOI: 10.54254/2755-2721/2026.37251

Figure 2 demonstrates that GAN generation is significantly faster than diffusion models, and DDIM exhibits superior speed compared to DDPM. Therefore, to improve generative efficiency, it is essential to leverage the inherent speed advantage of GANs.

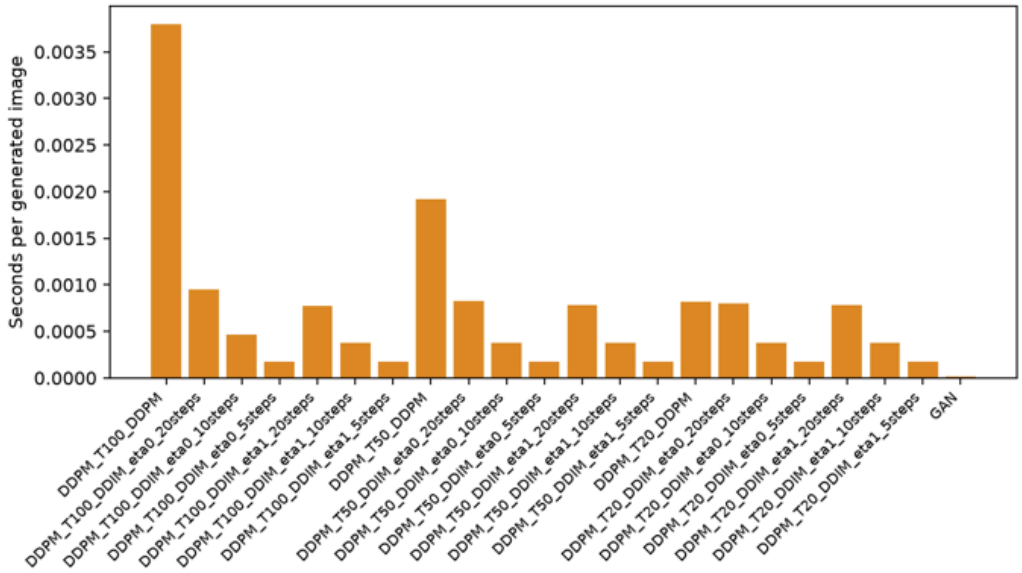

Figure 2. The measured sampling latency of different models (photo credit: original)

4.2. Fusion model

The DDIM function in Equation 9 selects the best-performing configuration, D1, to evaluate different hyperparameters λ and t_κ. Table 4 presents the performance of the fusion model under different hyperparameter combinations.

Table 4. Fusion model ablation study results (KID mean ± std over 3 random seeds)

KID	λ=0.1	λ=0.5	λ=1.0
t_κ=5	788.76±69.26	796.14±66.35	804.94±71.08
t_κ=10	737.54±67.51	745.02±68.39	752.54±66.22
t_κ=20	668.45±66.42	672.17±65.15	677.92±66.39

In the fusion model experiments, evaluate each configuration using three random seeds (42, 123, 456) and report the mean and standard deviation to validate the stability of hyperparameter selection. As shown in Table 4, the best KID performance is achieved by the model with λ=0.5, t_κ=20, yielding a KID of 672.17±65.15. The improvement is slight and falls within the standard deviation range, requiring further validation through additional repeated experiments. The generation quality of the fusion model is inferior to that of the baseline and D1, but superior to that of GAN. However, compared with D100, the fusion model reduces the Number of Function Evaluations (NFE) by 19%.

Figure 3 shows the loss variation of the GAN component in the fusion model, while Figure 4 shows the loss variation of the diffusion model component. Both curves exhibit oscillations, which may be attributed to an excessively large learning rate or the insignificant impact of the hyperparameter λ on the fusion model performance.

DOI: 10.54254/2755-2721/2026.37251

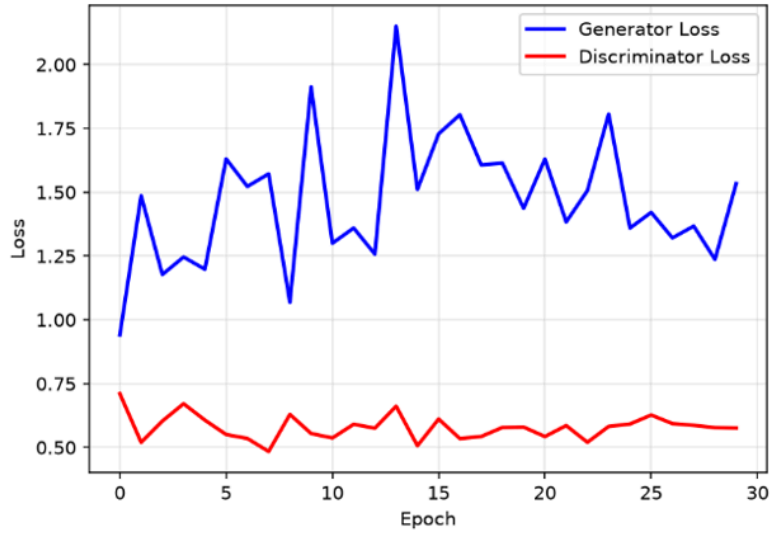

Figure 3. Training losses of GAN discriminator and generator over epochs (λ=0.5, t_κ=20) (photo credit: original)

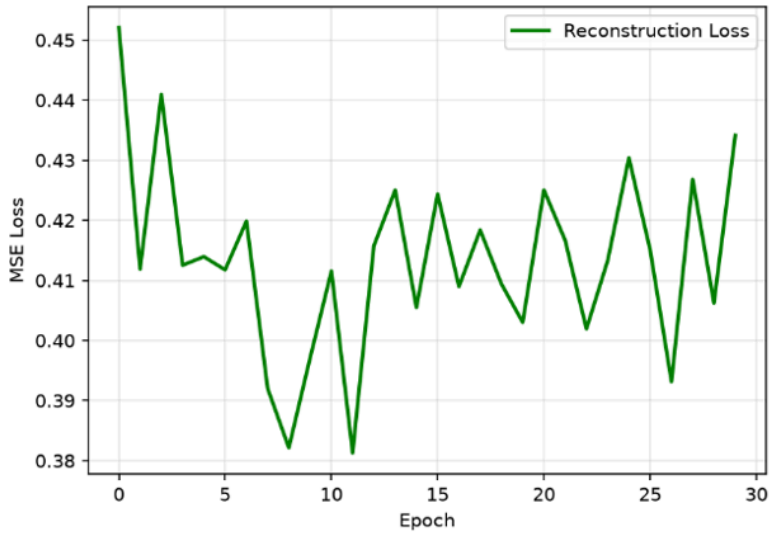

Figure 4. MSE loss over epochs (λ=0.5,t_κ=20) (photo credit: original)

5. Conclusion

This paper presents a systematic comparative study of GAN, DDPM, DDIM, and the fusion model on the Fashion-MNIST dataset. Experimental results show that diffusion models outperform GANs in terms of generation quality. DDIM with stochastic sampling achieves the best KID of 575.51±5.69 at 20 sampling steps, while deterministic sampling performs worst across all configurations. A shorter timestep (T=20) proves superior to T=100, indicating that simple data does not require a long denoising chain. Although the fusion model does not surpass pure diffusion models, it reduces the number of function evaluations to 81 steps, achieving approximately a 19% inference speedup compared to DDIM (100 steps).

Future work may treat the entire "GAN output + DDIM refinement" process as a unified generator and directly optimize the final output KID. Instead of using a fixed switching point t_K, Signal-to-Noise Ratio (log-SNR) can be used to determine the optimal transition step dynamically.

DOI: 10.54254/2755-2721/2026.37251

Future work could revisit the profiling-based timestep allocation strategy designed in this project. The profiling heuristic could compute the epsilon error, which is the expected squared L2 norm of the difference between the model's epsilon prediction and the true epsilon. For clipped predictions, finite differences in log-SNR could approximate the trajectory velocity and the curvature, which is the expected squared L2 norm of the divided difference of forward and backward velocities.

The fusion model adopts a serial structure where a conditional GAN first jumps from noise to an intermediate state, followed by DDIM refinement. However, this serial structure may not be the optimal fusion design. An alternative direction is to let GAN and diffusion models jointly determine the denoising direction at each step.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144.

[2] Zhu, R. (2023). Generative adversarial network and score-based generative model comparison. In *2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA)* (pp. 1–5). IEEE.

[3] Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.

[4] Lam, M. W. Y., Wang, J., Huang, R., Su, D., & Yu, D. (2021). Bilateral denoising diffusion models. arXiv preprint arXiv:2108.11514.

[5] Franceschi, J.-Y., Gartrell, M., Dos Santos, L., Issenhuth, T., de Bézenac, E., Chen, M., & Rakotomamonjy, A. (2023). Unifying GANs and score-based diffusion as generative particle models. arXiv preprint arXiv:2305.16150.

[6] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems, 33*, 6840–6851.

[7] Huang, Z., Wei, Y., & Chen, Y. (2026). Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *Mathematics of Operations Research*.

[8] Shen, Z., Mao, M., & Fan, P. (2024). A primary comparison of diffusion models and generative adversarial networks for image synthesis. In *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI 2024)* (pp. 225–234). Association for Computing Machinery.

[9] Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747.

[10] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456.

[11] Xu, Y., Tong, S., & Jaakkola, T. (2023). Stable target field for reduced variance score estimation in diffusion models. arXiv preprint arXiv:2302.00670.

[12] Li, Z., Wu, Y., & Chang, Y. (2026). AGGC: Adaptive group gradient clipping for stabilizing large language model training. arXiv preprint arXiv:2601.11864.