Improving trained diffusion model with adversarial loss for image generation

Yuval Gat and Boaz Shmueli

February 24, 2026

Abstract

Diffusion models have recently surpassed Generative Adversarial Networks (GANs) in image synthesis quality and distribution coverage. However, GANs remain highly effective at producing sharp, high-frequency visual details due to their adversarial training objective. In this project, we investigate whether adversarial fine-tuning can enhance a pretrained diffusion model by integrating it into a GAN framework. Specifically, we utilize the OpenAI Ablated Diffusion Model (ADM) as the generator and a pretrained ResNet18-based discriminator adapted for binary real/fake classification. After initializing the generator with diffusion weights, we jointly fine-tune the system under an adversarial objective. Our results reveal a clear sharpness–diversity trade-off: adversarial supervision significantly improves perceptual sharpness, but reduces precision and recall, indicating diminished distribution coverage. These findings highlight the fundamental tension between likelihood-based diffusion training and adversarial learning, and provide empirical insight into hybrid generative architectures.

Code is available at: https://github.com/11boaz11/Deep-Learning-final-project-Boaz-Yuval

1 Introduction

In recent years, generative models have experienced a significant leap in their ability to synthesize highly realistic images. Generative Adversarial Networks (GANs) have long dominated the field of image generation due to their capacity to produce sharp, high-resolution outputs. However, they frequently suffer from training instability and mode collapse. Recently, Diffusion Models have emerged as the new state-of-the-art, outperforming GANs in image synthesis benchmarks. While diffusion models offer unprecedented training stability and superior distribution coverage, their generated images may sometimes lack the pixel-level sharpness characteristic of adversarial networks. In this project, we propose a hybrid approach that synergizes the strengths of both paradigms by integrating a pre-trained diffusion model into a GAN architecture. The core idea is to utilize the trained diffusion model as the generator, and introduce a discriminator network tasked with distinguishing between real images and the diffusion-generated samples. By applying an adversarial objective to fine-tune the diffusion model, we aim to enhance image sharpness and investigate the complementary learning dynamics between diffusion-based generation and adversarial feedback.

2 Background

Our work is primarily inspired by the seminal paper *"Diffusion Models Beat GANs on Image Synthesis"* by Dhariwal and Nichol (2021) [1]. To contextualize our proposed hybrid architecture, we briefly review the theoretical foundations of both Diffusion Models and Generative Adversarial Networks (GANs).

2.1 Diffusion Models

Diffusion models are probabilistic generative models inspired by non-equilibrium thermodynamics. They operate through a two-step process:

- **Forward Process (Diffusion):** A Markov chain that systematically destroys the structure in a data distribution by gradually adding Gaussian noise to an image x_0 over T timesteps, eventually transforming it into pure isotropic Gaussian noise x_T.
- **Reverse Process (Denoising):** A neural network, typically a U-Net architecture with skip connections, is trained to reverse the forward process. It learns to progressively denoise x_t to reconstruct the original data x_0.

The training objective for the diffusion model is often simplified to predict the added noise ϵ at each timestep t:

$$\mathcal{L}_{diffusion} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{1}$$

The base paper also introduced the concept of *Classifier Guidance*, which utilizes gradients from a pre-trained classifier to guide the generation process toward specific classes, thereby improving image fidelity.

2.2 Generative Adversarial Networks (GANs)

GANs formulate image generation as a zero-sum game between two neural networks: a Generator (G) and a Discriminator (D). The Generator attempts to synthesize fake samples $G(z)$ from a latent noise vector z that perfectly mimic the real data distribution. Simultaneously, the Discriminator is trained to distinguish between real samples x and the generated fake samples.

The networks are trained jointly using the standard adversarial min-max objective:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2}$$

While GANs excel at producing sharp, high-fidelity images, they are notoriously difficult to train, often suffering from mode collapse and vanishing gradients. By utilizing a pre-trained diffusion model as the generator G in this framework, our project seeks to bypass the initial instability of GAN training while leveraging the discriminator's feedback to sharpen the diffusion outputs.

3 Contributions and Innovations

The primary contribution of our project is the novel integration of a pre-trained diffusion model within an adversarial learning framework. While standard diffusion models excel at data distribution coverage, they can sometimes lack the crisp, high-frequency details produced by GANs. By applying an adversarial objective to fine-tune a pre-trained diffusion model, we aimed to enhance image sharpness and overall visual fidelity. Our specific innovations and technical contributions are detailed below.

3.1 Hybrid Architecture Integration

We conceptualized and implemented a hybrid architecture where a pre-trained diffusion model (specifically, the OpenAI Ablated Diffusion Model - ADM) functions as the Generator within a GAN setup. For the Discriminator, we employed a Transfer Learning approach utilizing the classic ResNet18 architecture. Originally pre-trained to classify images into 1,000 distinct categories, we repurposed ResNet18 by modifying its final fully connected layers to output a single binary prediction. This adapted Discriminator is tasked with distinguishing real images from the fake samples generated by the diffusion process. The Generator was initialized with the pre-trained diffusion weights, and both components were jointly trained under a standard adversarial objective to refine the generated outputs.

3.2 Training Dynamics and Stability

Jointly training a diffusion process with an adversarial network introduces significant instability. To achieve stable convergence, we introduced several training techniques:

- **Discriminator Warm-up:** We allowed the Discriminator to train on a few initial batches before updating the Generator. This provided the Discriminator with a sensible baseline, preventing the Generator from receiving chaotic gradients early in training.
- **Curriculum Learning:** We structured the training complexity to increase gradually, allowing the model to achieve faster and more stable convergence.
- **Optimization Strategy:** We utilized the Adam optimizer, leveraging its normalized stochastic gradients with momentum to navigate the complex adversarial landscape.

3.3 Memory and Computational Optimizations

Fine-tuning a full-scale diffusion model alongside a discriminator is highly resource-intensive. To make this feasible, we implemented aggressive system optimizations:

- **Mixed Precision and Memory Management:** We utilized FP16 (half-precision) training to reduce VRAM footprint and applied aggressive memory cleanup protocols during the training loops.
- **Freezing Backbone Layers:** To save computational cycles and memory, we froze the deep backbone layers of the pre-trained diffusion model, focusing the gradient updates only on the layers most responsible for fine-grained image details.
- **Diffusion Timestep Respacing:** We applied timestep respacing during sampling to significantly accelerate the generation process without compromising the underlying Markov chain integrity.

3.4 Custom Evaluation Pipeline

To rigorously quantify the improvements achieved by our adversarial fine-tuning, we developed a custom evaluation suite. Beyond standard qualitative visual inspection, we implemented functions to compute exact **Sharpness** metrics, alongside **Precision and Recall** measurements, to evaluate whether the fine-tuned model successfully learned to generate sharper images while maintaining the original distribution coverage. We also utilized InceptionV3 feature extraction to validate the semantic integrity of the generated samples.

4 Results

To evaluate the impact of adversarial fine-tuning on the diffusion model, we employ three primary metrics:

- **Sharpness (Laplacian Variance):** Measures the presence of high-frequency details and edge definition. A higher variance indicates clearer, less blurry images.
- **Precision:** Quantifies the fidelity of the generated samples by measuring the extent to which they fall within the real data manifold.
- **Recall:** Evaluates the diversity of the model by measuring how much of the real data distribution is covered by the generated samples.

Table 1 presents the quantitative comparison between the pretrained diffusion baseline (ADM) and our adversarially fine-tuned hybrid model.

Metric	Pretrained	Fine-tuned	Change
Sharpness	0.0355	0.1385	+289.6%
Precision	0.4200	0.2800	-33.3%
Recall	0.1400	0.0300	-78.6%

Table 1: Quantitative comparison between the pretrained diffusion model and the adversarially fine-tuned version.

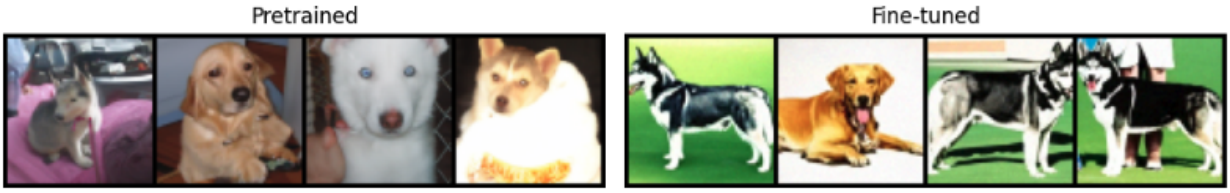

Figure 1: Sample generations from the pretrained diffusion model (left) and the adversarially fine-tuned model (right). The fine-tuned model exhibits sharper textures and stronger edge definition.

Adversarial fine-tuning led to a substantial improvement in image sharpness, with a relative increase of 289.6% compared to the pretrained diffusion baseline. This confirms that introducing a discriminator successfully enhances high-frequency details, edge definition, and local texture consistency in generated samples.

However, this perceptual improvement was accompanied by a degradation in distributional metrics. Precision decreased by 33.3%, indicating that the generated samples deviated more frequently from the true data manifold. Recall dropped sharply by 78.6%, suggesting a significant reduction in distribution coverage and potential mode concentration.

These results reveal a clear sharpness–diversity trade-off. While the adversarial objective strengthens perceptual realism, it partially compromises the diffusion model's inherent advantage in maintaining broad data support.

5 Conclusions and future directions.

In this project, we investigated whether adversarial fine-tuning can enhance a pretrained diffusion model by integrating it into a GAN framework. Our results demonstrate a clear trade-off: while adversarial supervision substantially improves perceptual sharpness, it simultaneously reduces precision and recall, indicating diminished distribution coverage.

The sharpness improvement confirms that discriminator-based feedback effectively enhances high-frequency details and local texture realism. However, the significant drop in recall suggests partial mode concentration, reflecting a reduced ability to preserve the full diversity of the data manifold. These findings highlight a fundamental tension between likelihood-based diffusion training and adversarial objectives.

Importantly, our empirical observations are consistent with the central insight of *"Diffusion Models Beat GANs on Image Synthesis"* (Dhariwal and Nichol, 2021). The original paper demonstrated that diffusion models outperform GANs largely due to their strong distribution coverage and training stability. By introducing adversarial fine-tuning, we partially reintroduce GAN-like behavior: sharper images, but weaker coverage. Thus, our results reinforce the theoretical distinction between probabilistic diffusion training and adversarial learning paradigms.

Future work could explore more balanced hybrid objectives, such as weighting the adversarial loss with a tunable coefficient, applying regularization techniques to stabilize the discriminator, or restricting fine-tuning to specific layers of the diffusion model. Additionally, evaluating the model with complementary metrics such as FID or human perceptual studies could provide further insight into the sharpness–diversity trade-off observed in this study.

Overall, our findings suggest that while adversarial fine-tuning can enhance perceptual quality, preserving the probabilistic structure learned by diffusion models remains a central challenge in hybrid generative architectures.

References

[1] P. Dhariwal and A. Nichol, *Diffusion Models Beat GANs on Image Synthesis*, Advances in Neural Information Processing Systems (NeurIPS), 2021.