Latent Denoising Diffusion Probabilistic Models

Team 11

Ashir Gowardhan
Carnegie Mellon University
agowardh@andrew.cmu.edu

Dhruv Gupta
Carnegie Mellon University
dhruvg2@andrew.cmu.edu

Yan Wang
Carnegie Mellon University
yanwang5@andrew.cmu.edu

1. Abstract

Diffusion models have emerged as a robust class of generative models, excelling in tasks like image synthesis, video generation, and text-to-image transformations. This project implements and evaluates Denoising Diffusion Probabilistic Models (DDPM) and Denoising Diffusion Implicit Models (DDIM), focusing on their ability to generate high-quality images. The input to these models includes image datasets such as CIFAR-10 (32×32 resolution) and ImageNet-100 (128×128 resolution), with outputs being generated images reconstructed from noise through iterative denoising processes. Key components of our implementation include the U-Net architecture for noise prediction, a DDPM noise scheduler for forward and reverse diffusion processes, and enhancements like Variational Autoencoders (VAE) to map high-dimensional data to latent space and Classifier-Free Guidance (CFG) for conditional image generation. The project evaluates model performance using metrics such as Frechet Inception Distance (FID) and Inception Score (IS), analyzing both the fidelity and diversity of the generated samples.

Diffusion models have significant applications in image synthesis, text-to-image generation, medical imaging (e.g., MRI enhancement), 3D object generation (e.g., gaming and VR assets) among others. Our initial results indicate that DDPM effectively models noise-injection and denoising processes, while DDIM significantly accelerates the inference process. Initial experiments revealed limitations, such as blurry image outputs and high computational costs. This study not only deepens understanding of diffusion model mechanisms but also lays a foundation for further innovations in image synthesis, text-to-image generation, medical imaging, etc.

2. Introduction

Diffusion models are a powerful class of generative models that have gained attention for their ability to generate high-quality and diverse data, such as images and videos. These models function by progressively adding random noise to the input data and then training a neural network to reverse the noise process step by step. This iterative denoising process allows the model to reconstruct meaningful outputs from pure noise. The framework has seen widespread application in domains such as image synthesis, text-to-image generation, and medical imaging.

In this project, we address the challenge of understanding and optimizing diffusion models for generative tasks, particularly image synthesis. The inputs to our models consist of image datasets such as CIFAR-10, a widely used dataset with 60,000 32×32 resolution images across 10 classes, and ImageNet-100, a larger dataset with 128×128 resolution images.

Our implementation focuses on Denoising Diffusion Probabilistic Models (DDPM) and Denoising Diffusion Implicit Models (DDIM), which serve as the backbone of our generative framework. The DDPM models the gradual noise injection and removal processes using a U-Net architecture, which is well-suited for capturing both local and global image features. DDIM introduces a non-Markovian approach to the reverse diffusion process, significantly accelerating inference without compromising output quality. To further enhance performance, we integrate Variational Autoencoders (VAE) for efficient latent space modeling and Classifier-Free Guidance (CFG) for conditional image generation, enabling greater control and improved image fidelity.

The motivation behind this study is the growing demand for generative models that can efficiently produce high-quality outputs across a range of applications, from image synthesis to medical imaging. By exploring the mechanisms of diffusion models and extending their capabilities through advanced techniques like noise scheduling and latent space optimization, we aim to contribute to the broader understanding and practical utility of these models. This project not only demonstrates the foundational strengths of DDPM and DDIM but also outlines a pathway for further innovation in generative modeling, bridging the gap between theoretical advancements and real-world applications.

3. Literature Review

Diffusion models have emerged as a highly effective approach for generative modeling, particularly in tasks like high-quality image synthesis and other data generation applications. Originally introduced for applications in 2D image generation, these models have since been adapted to broader tasks, leveraging a unique framework where noise is added to data through a diffusion process and reversed through a learned denoising model. Recent studies have explored the underlying mechanisms and potential improvements of Denoising Diffusion Probabilistic Models (DDPM), advancing the field of generative modeling significantly.

Previous research has focused on optimizing diffusion models to enhance both performance and efficiency. Ho et al. (2020) introduced the foundational concept of DDPM, showing its capability to generate high-resolution images by progressively removing noise through a Markov chain. The model's effectiveness in capturing complex data distributions has led to further advancements, such as Denoising Diffusion Implicit Models (DDIM), proposed by Song et al. (2020), which reduces the number of inference steps required for high-quality image synthesis. DDIM employs a deterministic alternative to traditional DDPM processes, achieving faster generation without sacrificing sample quality. This approach offers a significant improvement over the original model, especially for applications requiring efficient inference.

Further advancements have incorporated Variational Autoencoders (VAE) to improve the efficiency of diffusion models by compressing high-dimensional input data into a structured latent space. VAEs, as introduced by Kingma and Welling (2013), employ an encoder-decoder architecture to reduce data dimensionality while preserving essential features. In the context of diffusion models, using VAE-based latent spaces significantly lowers computational overhead, enabling faster and more scalable training while maintaining high image quality. This approach has proven to be particularly impactful for high-resolution image synthesis tasks, where working directly in pixel space is computationally expensive.

Recent studies have also explored the integration of Classifier-Free Guidance (CFG) for conditional image generation, addressing the limitations of earlier classifier-based approaches. Introduced as a simplified and effective method, CFG leverages conditional and unconditional training within a single model, eliminating the need for a separate classifier. By interpolating between the conditional and unconditional score functions, CFG enables flexible control over image generation, improving sample quality while simplifying the overall framework. This technique has demonstrated effectiveness across various generative tasks, providing a robust solution for conditional generation without compromising efficiency or simplicity.

Subsequent research has focused on improving both the performance and efficiency of diffusion models. Nichol and Dhariwal's "Improved Denoising Diffusion Probabilistic Models" introduced refined noise scheduling and learned variances, leading to smoother optimization and the ability to generate high-quality samples with fewer diffusion steps. This refinement demonstrated the critical role of noise schedules, prompting further investigation into how variance and scheduling strategies influence sample quality and generation speed. For our experimentation, we tested with three types of noise scheduling, namely Sigmoid, Cosine and Scaled Linear. This project builds upon these foundational models and techniques, implementing and evaluating DDPM, DDIM, VAE, and CFG to gain deeper insights into their mechanisms and contributions to generative modeling.

4. Model Description and Baseline

Our task is aiming to gain deeper insights into DDPM and DDIM mechanisms through experiments and implementation. Additionally, we extended the work by exploring Variational Autoencoders (VAE) and Classifier-Free Guidance (CFG) to further improve image generation quality.

4.1 DDPM and DDIM Baseline Description

The first baseline involves implementing a **Denoising Diffusion Probabilistic Model (DDPM)** from scratch. This model serves as the foundational approach for understanding the noise-injection and denoising process central to diffusion models. By training DDPM on progressively noisy data, we evaluate its ability to reconstruct clear and realistic images through step-by-step noise removal. Performance metrics like Frechet Inception Distance (FID) and Inception Score (IS) will be used to

quantify the quality and diversity of generated samples. This baseline provides a clear reference point for assessing improvements in image fidelity and generation robustness.

The DDPM model serves as the foundational diffusion model in our baseline, capturing the core mechanics of generative processes through progressive noise addition and removal. Formally, the forward process gradually corrupts the data x0 by adding Gaussian noise across multiple time steps t, modeled as a Markov chain. The reverse process, parameterized by pθ, seeks to denoise the data step-by-step, approximating the original data distribution by learning to remove noise at each time step.

The second baseline explores the **Denoising Diffusion Implicit Model (DDIM)**, which introduces a deterministic approach to reduce the number of sampling steps during inference. DDIM achieves faster sampling while retaining high image quality, making it a more efficient alternative to the traditional DDPM. By comparing DDIM to DDPM, we aim to demonstrate the efficiency gains in inference speed and analyze any trade-offs in quality. This comparison highlights the practical benefits of DDIM for applications where faster generation is essential.

The DDIM model introduces a deterministic approach to accelerate sampling, allowing fewer steps to achieve similar quality as DDPM. This is achieved by formulating a non-Markovian process in the reverse direction, which eliminates the randomness in the reverse process and thereby reduces the number of required steps.

4.2 Enhancing Implementation: CFG and VAE

Building on our work with DDPM and DDIM, we further extended our experiments by implementing Classifier-Free Guidance (CFG) and Variational Autoencoders (VAE) to improve image generation quality.

CFG was incorporated to enhance conditional image generation by integrating conditional and unconditional training directly into the diffusion model, eliminating the need for external classifiers. This approach simplifies the implementation and provides greater control over the generation process through the use of guidance weights. We tested various guidance scale values ranging from 3 to 6 to optimize the balance between flexibility and accuracy in the generated outputs.

Similarly, we implemented **VAE**s to create a structured and efficient latent space for diffusion processes. The VAE encoder compresses high-dimensional data into a lower-dimensional latent representation, while the decoder reconstructs the original data. This latent space significantly reduces the computational cost of the diffusion process and ensures that the essential features of the data are captured in a Gaussian-distributed form. By incorporating VAEs, we achieved faster and more effective high-resolution image generation, extending the capabilities of our diffusion models.

4.3 U-Net Architecture

The U-Net architecture serves as the core component for the noise prediction network within our diffusion models. Originating from biomedical image segmentation, U-Net has gained widespread adoption in

generative modeling tasks due to its capacity to capture both low-level details and high-level contextual information. Its encoder-decoder structure with skip connections makes it particularly effective for iterative refinement tasks like denoising. The U-Net is composed of a contracting (encoder) path and an expanding (decoder) path. The encoder progressively downsamples the spatial resolution of the input while increasing the feature dimensionality, extracting hierarchical representations. The decoder then symmetrically upsamples and refines these representations back to the original input resolution. Skip connections link corresponding layers in the encoder and decoder, ensuring that spatial information lost during downsampling is reintroduced. These connections help preserve fine-grained details and stabilize training, allowing the model to produce sharper, more coherent reconstructions. In diffusion models, a time embedding is integrated into the U-Net to condition its output on the diffusion step. This embedding provides a temporal context that guides the denoising process at each iteration. For conditional image generation scenarios, such as using Classifier-Free Guidance (CFG), additional conditioning information (e.g., class or text embeddings) can be injected into the U-Net's feature maps, making it a flexible backbone for diverse tasks. To handle complex data distributions and incorporate external conditioning (such as text prompts or class labels), attention mechanisms may be integrated into the U-Net. Standard self-attention layers can capture long-range dependencies within the feature maps, while cross-attention blocks allow the model to relate external conditions (e.g., text tokens) to image features effectively.

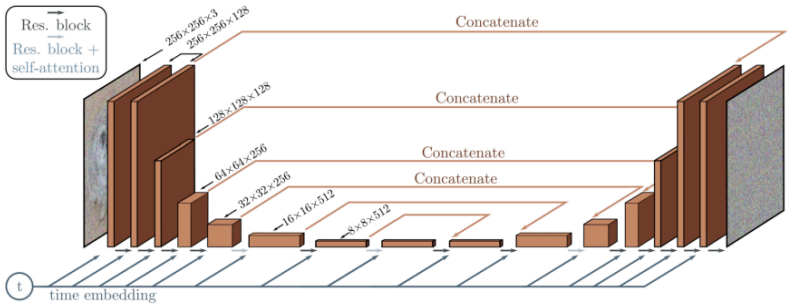


Fig: DDPM architecture (referenced from HW5 handout)

5. Dataset

Initially, we began our experiments with the CIFAR-10 dataset, consisting of 60,000 images (50,000 for training and 10,000 for testing) across 10 classes (e.g., airplane, car, bird, etc.). Each image originally had a resolution of 32×32 pixels, making it a computationally efficient choice for testing our model architecture. We applied transformations such as resizing the images to 128×128 pixels and normalizing the pixel values to the range [-1, 1]. This smaller dataset allowed us to observe results quickly and troubleshoot any issues without the high computational demands of larger datasets.

As our project progressed, we transitioned to the ImageNet-100 dataset to scale our experiments and test our models on more complex and diverse data. This dataset consists of 126,689 training images and 5,000 validation images spanning 100 classes (e.g., goldfish, great white shark, etc.), with each image resized to 128×128 pixels and normalized to the range [-1, 1]. Using ImageNet-100 enabled us to evaluate the

robustness and scalability of our models, providing a more rigorous assessment of their performance in handling high-resolution and diverse image distributions.

Dataset Summary	Experiment Phase 1	Experiment Phase 2
Dataset Used	CIFAR-10	ImageNet-100
Dataset Size	60,000 images	131,689 images
Training Images	50,000	126,689
Testing/Validation Images	10,000 (testing images)	5,000 (validation images)
Image Resolution	32×32 pixels	128×128 pixels
Number of Classes	10 (e.g., airplane, car, bird, etc.)	100 (e.g., goldfish, great white shark, etc.)
Data Transformation	Resizing to 128×128; Normalized to [-1,1]	Resizing to 128×128; Normalized to [-1,1]

Overview of Datasets Used in Experiments

6. Loss functions

The DDPM loss function, often referred to as L□ᵢ□□□ₑ, is essentially a mean squared error (MSE) between the predicted noise and the actual noise added to the image15. The loss function can be expressed as:

$$L_{\text{simple}} = E_{t,x_0,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right]$$

Some other losses that could have been used were:

- Variational Lower Bound (VLB) Loss: Also known as Evidence Lower Bound (ELBO) Loss, it's used in Denoising Diffusion Probabilistic Models (DDPMs) to approximate the true posterior distribution of the data given the noise.
- L1 Loss (Mean Absolute Error): Measures the average absolute difference between predicted and actual pixel values.
- L2 Loss (Mean Squared Error): Measures the average squared difference between predicted and actual pixel values.
- Perceptual Loss: Combines content loss and style loss to capture high-level features and textures, using pre-trained networks like VGG.

7. Evaluation Metrics

We used two metrics in the evaluation of generative models are the *Frechet Inception Distance (FID)* and the *Inception Score (IS)*. These metrics provide quantitative measures to assess the quality and diversity of the generated data, allowing us to compare models more systematically.

- **Frechet Inception Distance (FID)**

  It measures the Frechet distance between these distributions in a feature space derived from the Inception network. Lower FID scores indicate a closer alignment between the distributions of real and generated data, reflecting better model performance.

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

- **Inception Score (IS)**

  The score takes into account how confidently the model assigns labels to generated samples (indicating quality) and whether those samples cover a wide range of diverse classes (indicating diversity). Higher IS values correspond to higher quality and diversity of generated samples.

$$\text{IS} = \exp\left(\mathbb{E}_x\left[D_{\text{KL}}(p(y|x)||p(y))\right]\right)$$

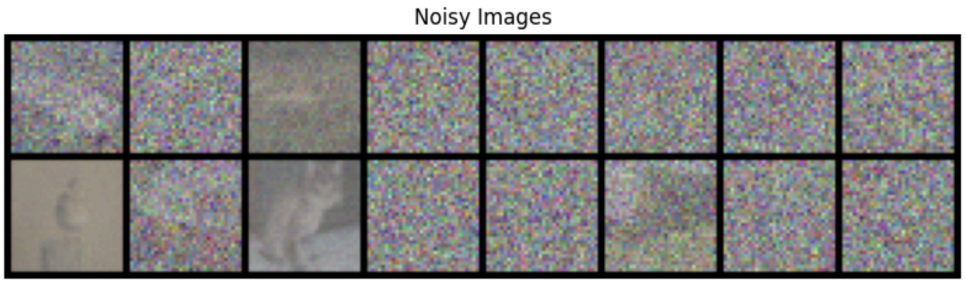
8. Results

8.1 Qualitative Results

Initially, we attempted to use the ImageNet-100 dataset for training; however, the computation time required for training on this dataset proved to be significantly high. To establish a more manageable starting point, we transitioned to the CIFAR-10 dataset. The initial images and corresponding noisy images used for training are shown below:

```
Displaying initial batch of original images:
```


```
Displaying initial batch of noisy images:
```



In the next phase, we trained our DDPM model with 1000 training time steps and 1000 inference steps. After further refinements to improve computational efficiency, we reduced the configuration to 200 training time steps and 200 inference steps. Subsequently, we trained the DDIM model using 1000 training time steps and 200 inference steps. The images generated during DDPM training are displayed below:

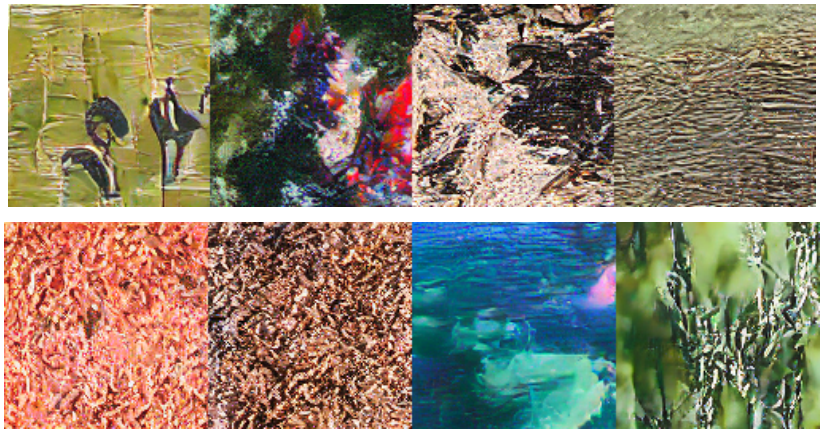
Figure: ImageNet

The initial experiments revealed that the generated images were quite blurry, suggesting the need for further training with improved parameters. Building on this, we conducted training with DDIM, using 1000 training time steps and 200 inference steps. The images generated during DDIM training are shown below:

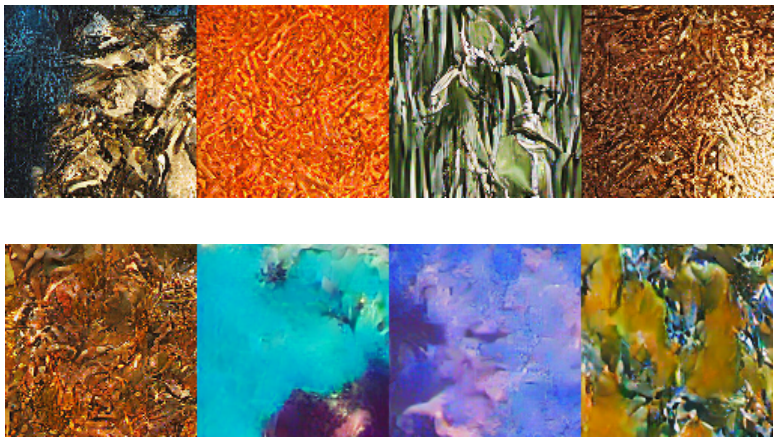
Figure: ImageNet

To further enhance the model's capabilities, we implemented Variational Autoencoders (VAE) and Classifier-Free Guidance (CFG). Although we encountered scaling issues during the VAE experiments, adjustments were made to resolve these challenges. The images generated during VAE training are depicted below:

Subsequently, we trained the model with CFG, which improved the generation quality and provided more flexibility in the image generation process. The images produced during CFG training are presented below:



Finally, we combined VAE with CFG in our DDPM model, achieving the best results among all configurations. The images generated during the training of DDPM with VAE and CFG are displayed below:



8.2 Code and Ablations Repository

Github Project Code Link: Code Link

Kindly request access to view, for privacy and security reasons we have kept the access to our code request-only. It is not public.

Wandb Ablations Link: Wandb ablations link

The link contains the images, our saved models, and the info for our runs.

9. Inference

We also ran the inference by loading the checkpoints we got from our training from both DDPM and DDIM. Due to limited time and computational availability, we only tested our FID and Inception Score using only 10 images for Cifar. For image-net we proceeded to run it on 200 images. The number of images were selected due to limitations in compute and time.

Model	DDPM	DDIM	VAE	DDPM+CFG	DDPM+VAE+CFG
FID	519 on Cifar 340 on Image-Net	- 556 on Cifar - 335 on Image-Net	332	290	248.5
Inception Score	1.2757, 1..0749	1.4757, 1.0449	1.357, 1.72	1.33, 1.021	1.0849, 1.0610

Performance Comparison of Models Across Metrics (FID and Inception Score)

10. Exploring More Advanced Training Techniques

Beyond the foundational implementation of DDPM and DDIM in this project, there are several advanced techniques in diffusion models that have demonstrated significant improvements in performance and efficiency.

10.1 Advanced Noise Scheduling Strategies

The linear noise schedule used in standard DDPMs is straightforward but often suboptimal for complex datasets. Techniques such as cosine noise scheduling redistribute noise more effectively across time steps, improving both generation speed and sample fidelity. These advanced schedules optimize the training process and are particularly impactful when scaling models to larger datasets. For noise scheduling we experimented with three different scheduling strategies, namely sigmoid, cosine and scaled Linear. The links for the following can be found in the links attached sigmoid and cosine.

10.2 Learnable Variance in Gaussian Transitions

Traditional DDPMs rely on fixed variance for stability, but introducing learnable variance offers greater flexibility. This approach allows the model to dynamically adapt to data distributions, enhancing image quality and aligning diffusion processes closer to Variational Autoencoder (VAE) principles. This innovation has shown the potential to generate more diverse and accurate samples. As far as the implementation of learned variance is concerned, we made an attempt to implement a basic version for the same but were unable to get it to work and produce appropriate results.

10.3 Model Architecture Enhancements

Enhancing the U-Net backbone, such as increasing its size or incorporating more advanced architectures like Diffusion Transformers (DiT), allows diffusion models to better handle high-resolution images and intricate details. These architectural innovations offer a pathway to improve the capacity and versatility of diffusion models. While we did explore coding this into our project, due to time constraints we were not able to get it working, and have excluded it from the project as a result.

This paper proposes diffusion models based on the Transformer architecture, demonstrating their advantages in scalability and performance.

11. Conclusion

In this project, we successfully implemented and experimented with multiple configurations of diffusion models, beginning with DDPM and DDIM. Our initial training on the CIFAR-10 dataset allowed us to test and refine these models, providing a solid foundation for further exploration. Building on this, we implemented Variational Autoencoders (VAEs) and Classifier-Free Guidance (CFG) to enhance image generation quality. Despite facing scaling issues during VAE experiments, we identified the problem and proceeded effectively.

Our experiments revealed that combining DDPM with CFG achieved decent results, while the combination of DDPM with VAE and CFG delivered the best performance. To further optimize the models, we explored different noise scheduling methods, including cosine, sigmoid, and scaled linear schedules. Among these, the scaled linear noise schedule (specific to VAE) produced the most promising results, whereas cosine scheduling did not perform well. Unfortunately, due to time and computational

constraints, we were unable to run inference for some configurations or fully explore advanced techniques like learned variance and Diffusion Transformers (DiT).

Overall, the integration of DDPM with VAE and CFG proved to be the most effective approach in our experiments. While certain advanced techniques remain unexplored, our findings highlight the potential of diffusion models for high-quality image generation and lay a strong foundation for future work. Looking ahead, we aim to address computational constraints, refine noise scheduling strategies, and further investigate advanced methodologies to enhance model performance and scalability.

References

1. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems, 33*, 6840–6851.
2. Song, J., Meng, C., & Ermon, S. (2020). Denoising Diffusion Implicit Models. *arXiv preprint arXiv:2010.02502*.
3. Nichol, A., & Dhariwal, P. (2021). Improved Denoising Diffusion Probabilistic Models. arXiv preprint arXiv:2102.09672.
4. Song, Y., & Ermon, S. (2019). Generative Modeling by Estimating Gradients of the Data Distribution. Advances in Neural Information Processing Systems, 32.
5. Chen, T. (2023). On the Importance of Noise Scheduling for Diffusion Models. arXiv preprint arXiv:2301.11093. (From provided text; if exact citation unknown, use year and title as given.)
6. Hugging Face. (2023). Diffusers: State-of-the-art diffusion models for image and audio generation. Retrieved from https://huggingface.co/docs/diffusers

Division of Work: Our team equally divided the work amongst ourselves. Ashir was tasked with DDPM pipeline and scheduler implementation, Dhruv was tasked with DDIM scheduler and train and inference implementations. Yan was tasked with literature review, exploring solutions, code reviews and analysis. Ashir then worked on vae implementation, dhruv on cfg, and Yan explored different noise scheduling techniques. Each member contributed equally to the progress.