Lab 09 Report

This project introduced me to the fundamental concepts behind diffusion-based generative models and how they can progressively transform random noise into meaningful images. Through implementing, training, and evaluating a conditioned U-Net model on the MNIST dataset, I gained a clearer understanding of the delicate balance between randomness, structure, and learning in artificial intelligence.

In the forward diffusion process, clean images are gradually corrupted with small amounts of Gaussian noise over many steps. Each step slightly reduces the clarity of the image until it becomes nearly indistinguishable from random noise. By observing the noise progression visualization, I learned that gradual corruption is essential. Adding noise step-by-step allows the model to learn how to denoise at various stages instead of facing the impossible task of restoring a completely random image in one step. In my experiments, recognizable structure in digits like “1” or “7” began to reappear roughly halfway through the denoising process, while more complex digits such as “8” required more steps before becoming visible again. This shows that recognition depends both on the inherent complexity of the image and the model’s stage of reconstruction.

The U-Net architecture was particularly effective for this task because it combines global understanding with fine detail preservation. The encoder compresses the input into a latent representation, while the decoder reconstructs the image step-by-step, using skip connections to carry over important local details lost during downsampling. These skip connections proved vital, as they allowed the model to maintain edges and shapes that define the digit’s structure. The time embedding, encoded using sinusoidal functions, gave the model a sense of “position” in the diffusion timeline—essentially telling it how noisy the current image

was—while the class embedding enabled it to generate specific digits by conditioning on class labels.

During training, the mean squared error loss represented how accurately the model could predict the added noise. Early in training, the loss was high, and generated images were indistinct. Over time, as the loss decreased, the model produced progressively clearer and more realistic digits. This improvement highlighted how learning in diffusion models is not about memorizing images but understanding the dynamics of noise removal. It was also evident that training stability depended heavily on the learning rate and the number of diffusion steps, which influenced how smoothly the model converged.

Integrating CLIP for evaluation provided an objective method to measure image quality. CLIP compared generated samples to textual descriptions like “a handwritten number 7” or “a blurry number,” producing similarity scores that reflected the recognizability and clarity of each output. Simpler digits such as “1” and “4” consistently received higher CLIP scores, while complex digits with closed loops, like “8” or “9,” were more challenging to reconstruct convincingly. This result suggested that the model found it easier to represent features defined by straight or singular strokes. CLIP thus acted as an automated critic, offering quantitative insight into generative quality. In future iterations, CLIP feedback could be incorporated into training as an auxiliary objective, encouraging the model to produce samples that align more closely with human-like perception.

Beyond these evaluations, the project revealed both the potential and the limitations of current diffusion models. The main advantages lie in their flexibility and ability to produce diverse, high-quality samples. However, they remain computationally expensive and slow

due to the many denoising steps required for each image. Moreover, our model was limited to small grayscale images and lacked the ability to generate stylistic variations or higher-resolution outputs. If I were to continue this project, I would first implement a cosine noise schedule to improve training stability, then explore classifier-free guidance to achieve a better trade-off between fidelity and diversity, and finally adopt a faster sampler such as DDIM or DPM-Solver to reduce inference time.

Overall, this project was a meaningful exploration into modern generative modeling. The process of observing noise turn into structured digits felt both technical and philosophical—it demonstrated how order can emerge from randomness when guided by well-designed systems. The integration of CLIP connected visual and linguistic understanding, illustrating the growing convergence of multimodal AI. This journey deepened my appreciation for how generative models can learn not just to reproduce data, but to interpret and recreate patterns of meaning, bridging perception and creativity in artificial intelligence.