FashionMNIST Stable Diffusion: Technical Report

1. Executive Summary

The **FashionMNIST Stable Diffusion** project implements a state-of-the-art Latent Diffusion Model (LDM) tailored for the FashionMNIST dataset. By compressing high-dimensional image data into a lower-dimensional latent space, the system efficiently generates high-quality, synthetic apparel images. This implementation serves as a robust demonstration of generative modeling, showcasing the synergy between Variational Autoencoders (VAEs) and UNet-based denoising networks.

2. Project Objectives

The core mission of this project is to:

- **Implement Latent Diffusion**: Develop a complete pipeline for latent diffusion, moving beyond pixel-space diffusion for improved computational efficiency.
- **Generative Excellence**: Produce realistic, novel images that capture the stylistic nuances of the FashionMNIST apparel categories.
- **Modular Architecture**: Provide a clean, modular codebase in PyTorch that facilitates experimentation with different model components and hyperparameters.

3. Technical Architecture

3.1 Variational Autoencoder (VAE)

The VAE is responsible for the transition between image space and latent space.

- **Encoder**: Compresses 28x28 grayscale images into a 7x7 latent representation.

- **Decoder**: Reconstructs the image from the latent vector.
- **Key Features**: Utilizes ChannelAttention, GroupNorm, and SiLU activations to maintain high fidelity during compression and reconstruction.

3.2 UNet Denoiser

The UNet serves as the "epsilon model," predicting the noise added to the latent representation at various timesteps.

- **Structure**: Features a symmetric encoder-decoder architecture with skip connections.
- **Components**: Integrates CrossAttention and LatentResBlocks to effectively capture both local and global features within the latent space.

3.3 Diffusion Process

The system employs a Denoising Diffusion Probabilistic Model (DDPM) framework:

- **Forward Process**: Iteratively adds Gaussian noise to the latent vector over 1000 timesteps.
- **Reverse Process**: The trained UNet iteratively removes noise to recover the original latent distribution from pure noise.

4. Training and Configuration

The models were trained using the following standardized parameters:

Parameter	Value
Dataset	FashionMNIST (60,000 training samples)
Input Resolution	28 x 28 (Grayscale)
Latent Resolution	7 x 7
Diffusion Steps	1000
Training Epochs	100
Batch Size	128
Optimizer	Adam (Learning Rate: 1e-4)

5. System Workflow

1. **VAE Pre-training**: The VAE is trained to minimize reconstruction loss and KL divergence, ensuring a meaningful latent space.
2. **Latent Encoding**: Training images are passed through the VAE encoder to generate latent vectors.
3. **Noise Injection**: Random noise is added to the latent vectors according to a predefined variance schedule.
4. **UNet Training**: The UNet is trained to predict the added noise, conditioned on the timestep.
5. **Inference (Sampling)**: Starting from random noise, the UNet and VAE decoder work together to generate new apparel images.

6. Conclusion and Future Work

This project successfully demonstrates the power of Latent Diffusion Models on structured image datasets. The modular implementation provides a solid foundation for future enhancements, such as:

- **Conditional Generation**: Incorporating class labels to allow for targeted generation of specific apparel types (e.g., “Generate a sneaker”).

- **Higher Resolution**: Adapting the architecture for more complex datasets like CelebA or CIFAR-10.
- **Performance Optimization**: Exploring faster sampling techniques like DDIM to reduce inference time.

Report generated for the FashionMNIST Stable Diffusion Project