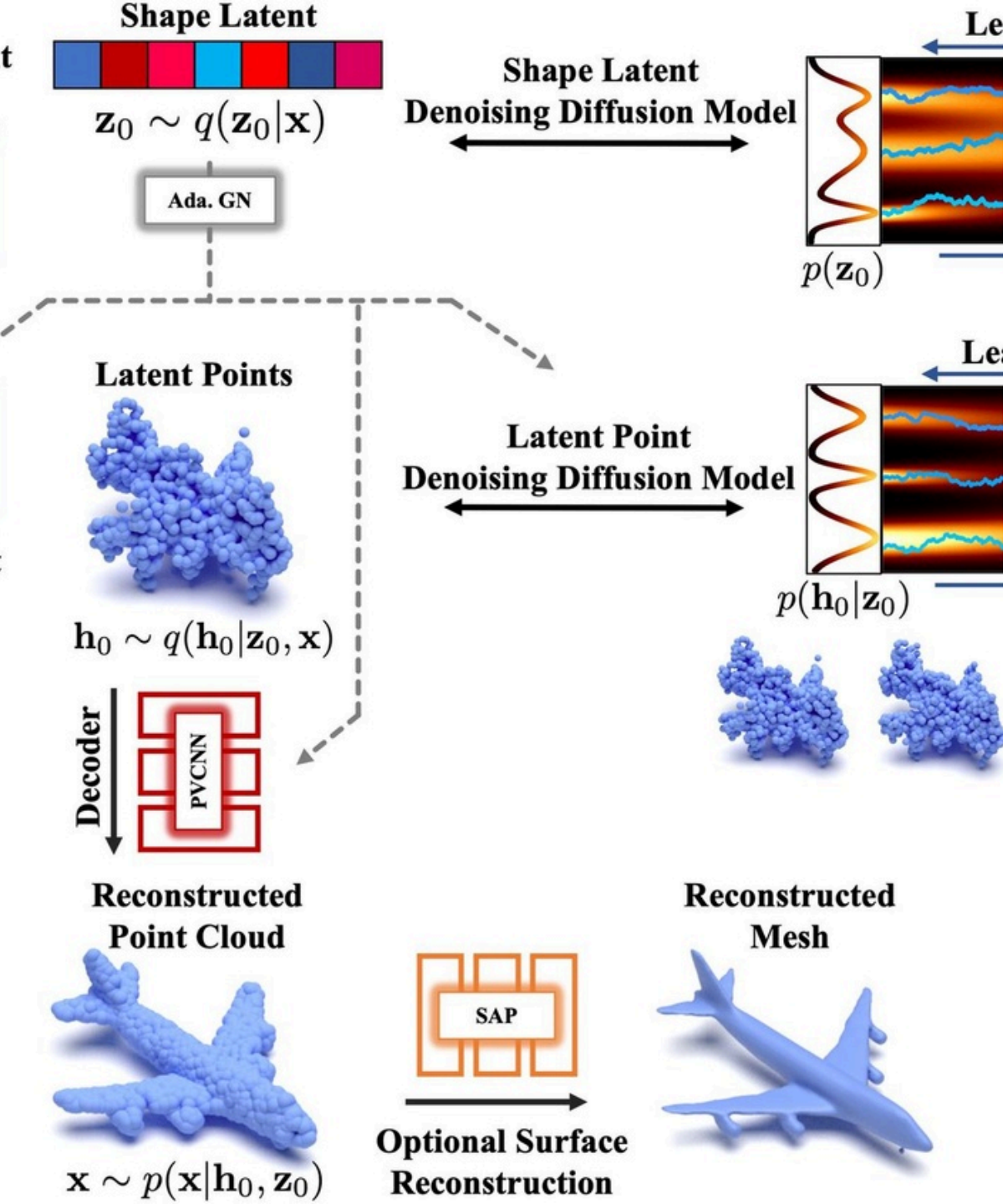
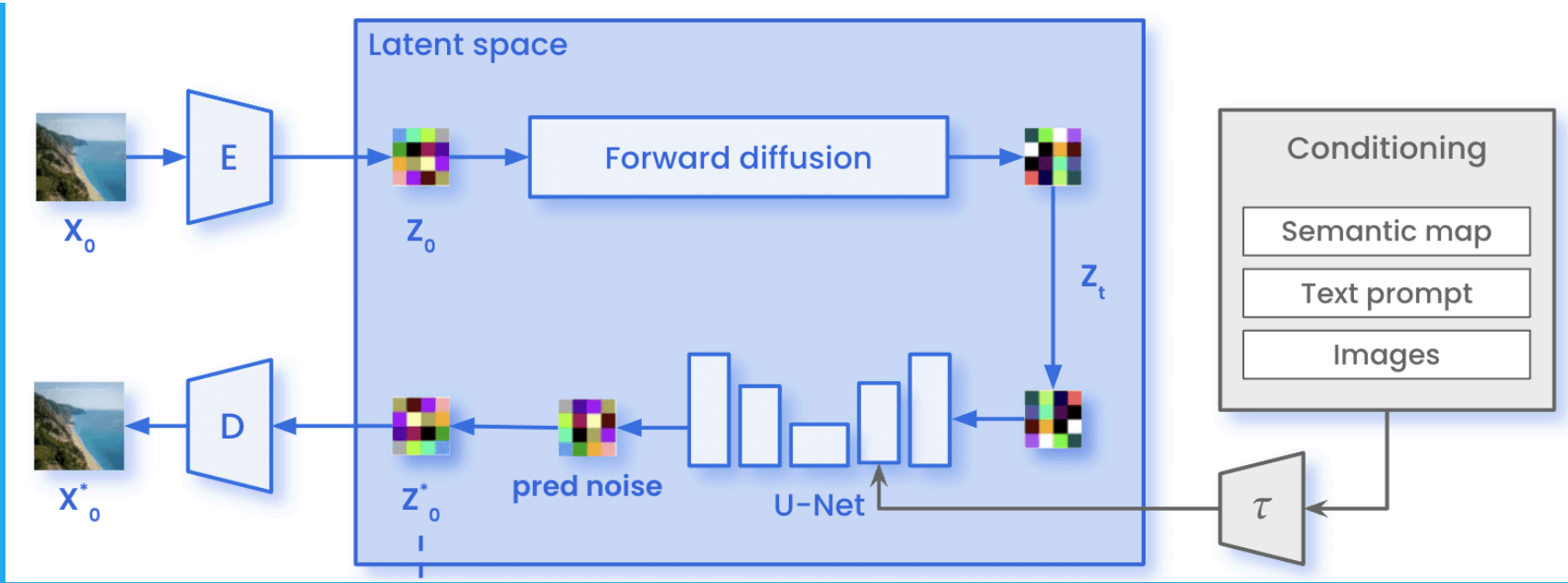


Latent Diffusion Model: A Full-Stack Application for Image Generation

A full-stack application for generating high-quality images from text prompts using CLIP tokenization, VAE encoding/decoding, and U-Net architecture.



CSYE 7380 FINAL PROJECT

GROUP - 1

DARSHIT SHARMA

MONISHA M

SIDDHARTH DUMBRE

Table of Contents

Problem Statement

The Challenge of Image Generation from Text

Generating high-quality images from text prompts is a challenging task that requires advanced natural language processing and computer vision techniques. Traditional methods often struggle to produce coherent and realistic images that match the given text descriptions.

Motivation for the Latent Diffusion Model

The Latent Diffusion Model project aims to address the limitations of existing image generation approaches by leveraging a combination of CLIP tokenization, VAE encoding/decoding, and U-Net architecture to generate visually compelling images from text prompts.

Potential Applications

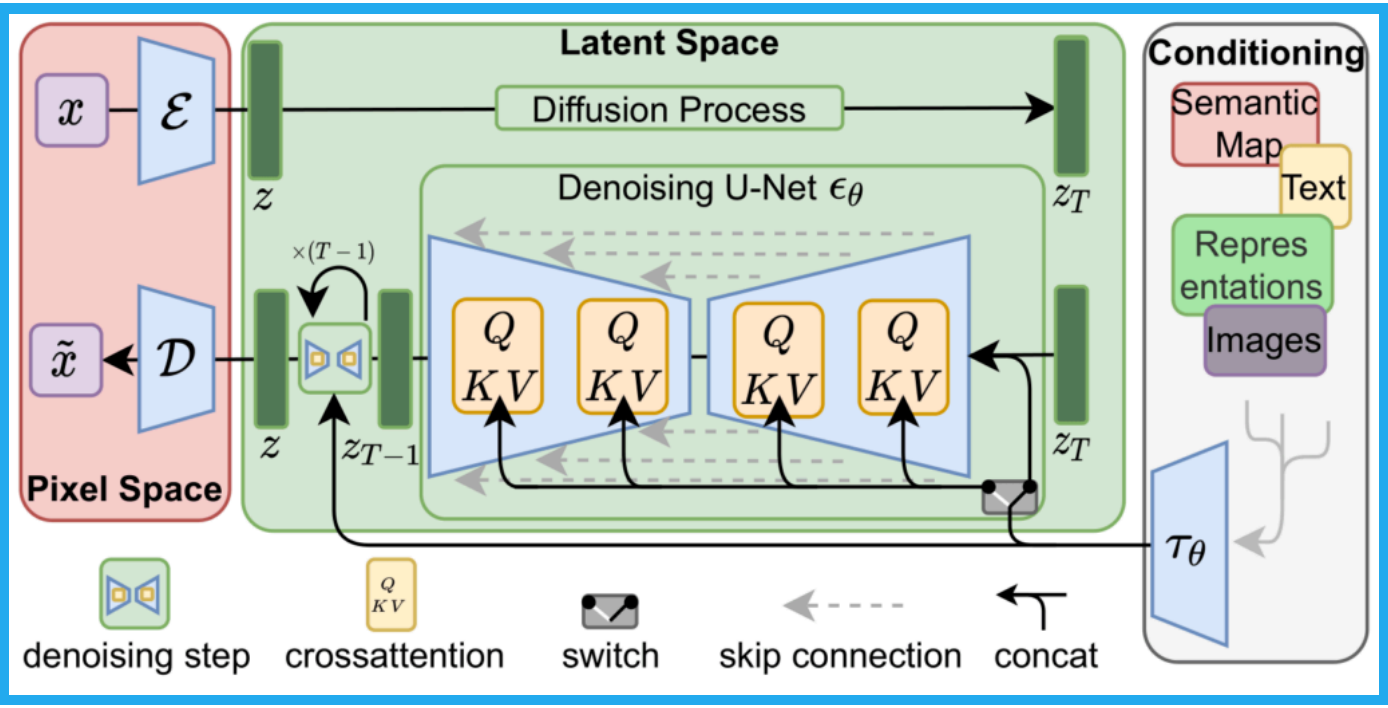
Successful development of the Latent Diffusion Model could enable a wide range of applications, such as content creation for digital media, interactive design tools, and personalized visual experiences, enhancing user engagement and creativity.

Improving Image Generation Quality

By fine-tuning the Latent Diffusion Model on the Naruto BLIP captions dataset, the team aims to push the boundaries of image generation quality, creating a more immersive and realistic experience for users.

OBJECTIVE

To develop a full-stack text-to-image Generative AI web application using Stable Diffusion, fine-tuned with a Naruto anime-style dataset, where only the U-Net weights are updated via **LoRA (Low-Rank Adaptation)** using **PEFT** for efficient parameter tuning. The system enables stylized image generation in anime form based on user prompts and is evaluated quantitatively using **CLIP Score** for semantic alignment. The application integrates a **FastAPI** backend with two separate APIs (base model & fine-tuned model), and a **Streamlit** frontend for user interaction.



Data Preparation

Acquiring the Naruto BLIP Captions Dataset

Name: lambdalabs/naruto-blip-captions (from Hugging Face)

Type: Anime images with BLIP-generated captions

Train set: All 1221 examples for fine-tuning

Image Preprocessing

Resize: All images resized to 256 x 256

Normalization: Pixel values scaled to [-1, 1]

Caption Tokenization

Tokenizer: CLIPTokenizer (from StableDiffusionPipeline)

Truncation & Padding: Enabled, with max_length = 77

Outputs: input_ids used as text conditioning

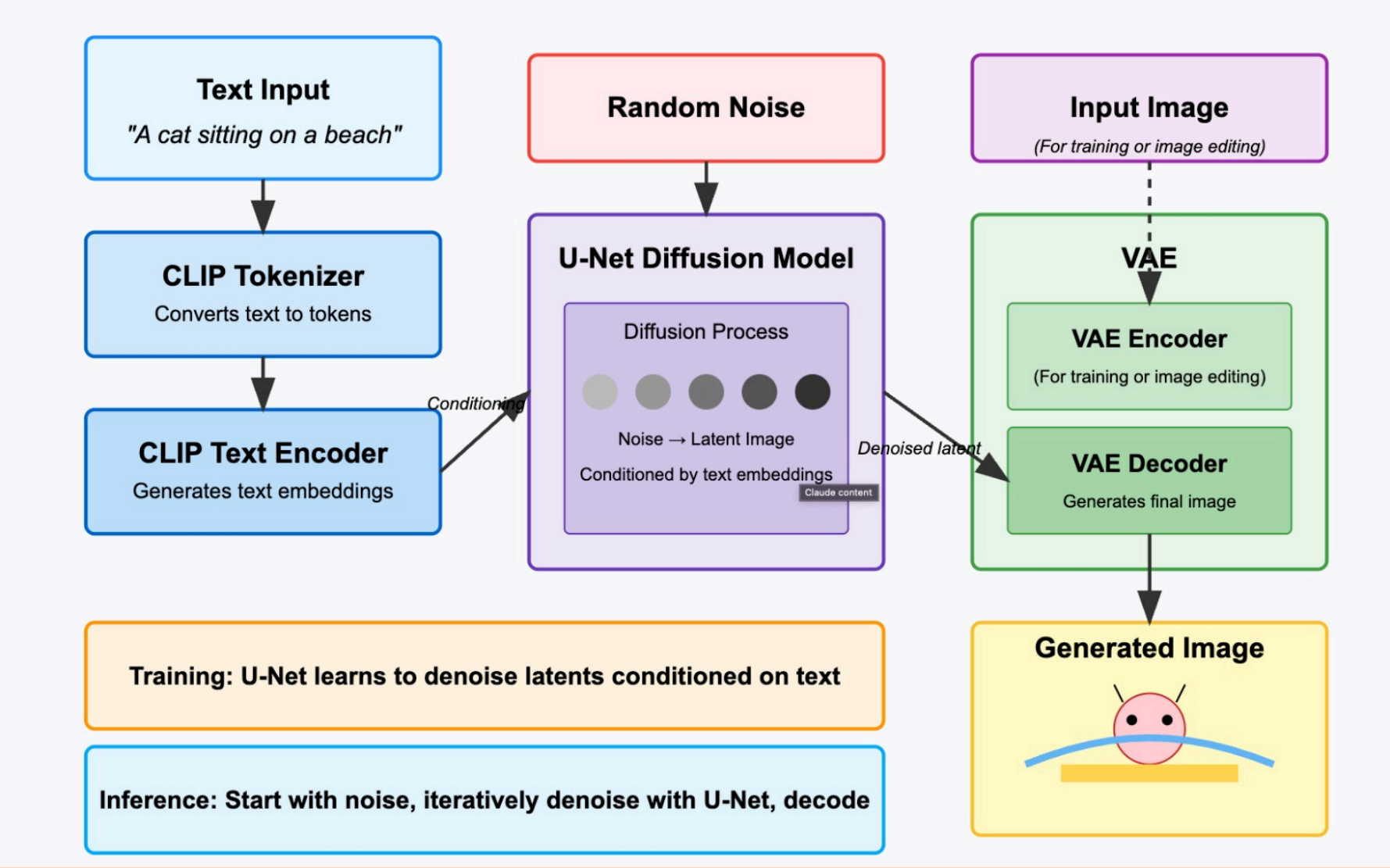
DataLoader Configuration

Batch Size: 2

Shuffling: Enabled for randomness during training

Framework: PyTorch torch.utils.data.DataLoader

Model Architecture Diagram



Model Architecture

1 CLIP Tokenizer

The CLIP tokenizer takes the text input and generates text embeddings, which capture the semantic and contextual information of the prompt.

2 VAE Encoder

The Variational Autoencoder (VAE) encoder converts the input images into their latent representations, which are low-dimensional and compressed versions of the original images.

3 VAE Decoder

The VAE decoder is responsible for reconstructing the images from their latent representations, effectively generating new images that are visually similar to the input.

4 U-Net Architecture

The U-Net architecture, which is the core of the Latent Diffusion Model, generates the final high-quality images by combining the text embeddings and the latent representations of the images.

5 Fine-tuning on Naruto BLIP Captions

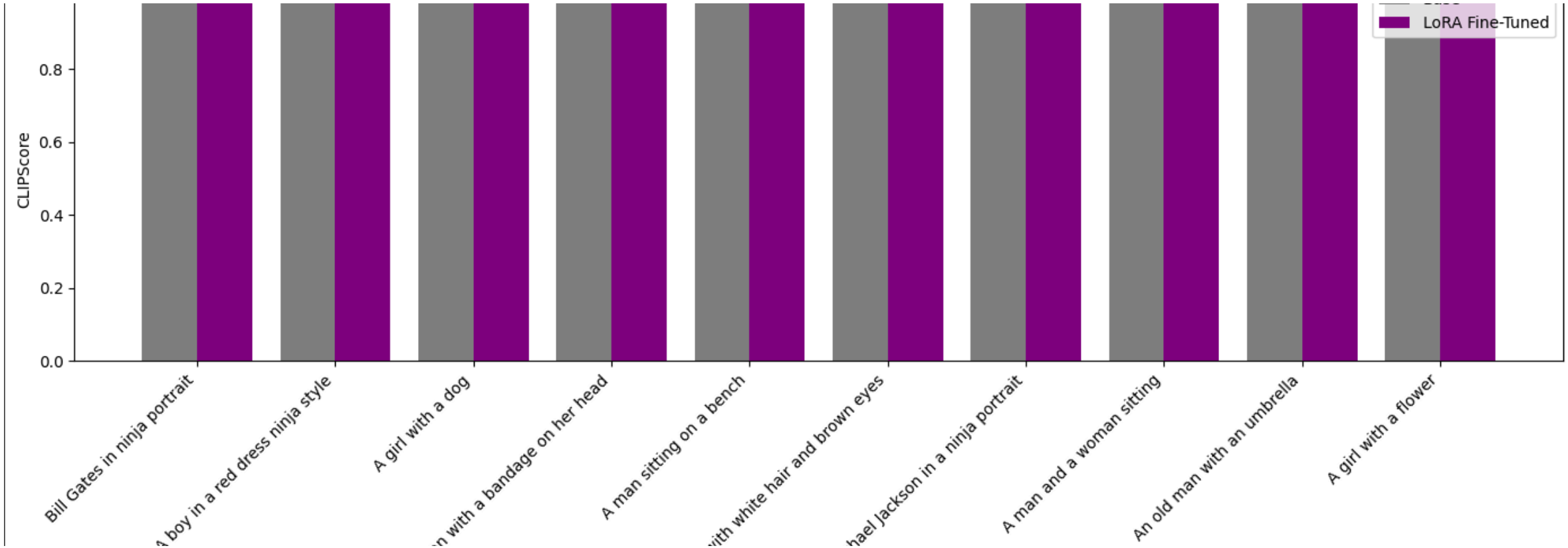
The Latent Diffusion Model has been fine-tuned on the Naruto BLIP captions dataset, allowing it to generate images that are closely related to the Naruto anime universe.

6 Inference

We perform inference after fine-tuning the u-net model in order to generate the anime images based on the given prompt

Model Evaluation

CLIP Score is a similarity metric that measures how well a generated image matches its text prompt using CLIP's joint image-text embeddings



*Calculated based on the model's performance on the Naruto BLIP captions dataset.

Generation Steps for ID: 71a7d2fe-1eda-493b-ae34-145a0b3b5b44

Step 10

Step 20

Step 30

Step 40

Step 49

Step 50

Results

A girl with a dog

Image generated on the base model

A girl with a dog

Image generated on fine tuned Model

A man wearing a hat

Image generated from base model

A man wearing a hat

Animated Image generated from the fine tuned model

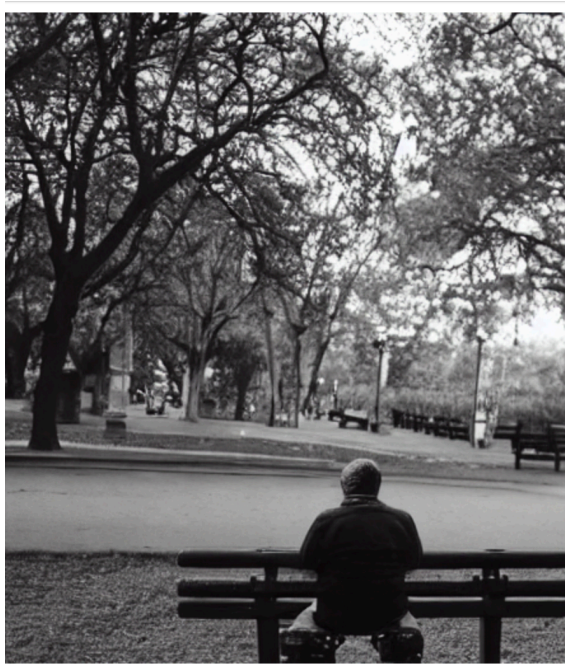
Results

Anime with red eyes and white hair

Image generated on the base model

Anime with red eyes and white hair

Image generated on fine tuned Model



A man sitting on a bench

Image generated from base model

A man sitting on a bench

Animated Image generated from the fine tuned model

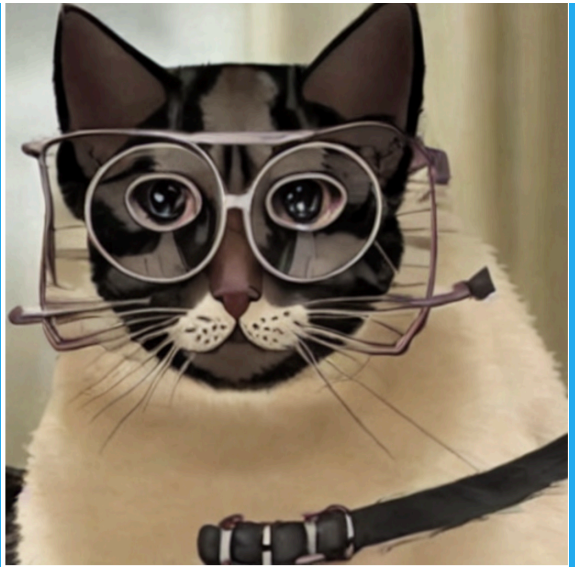
Observations

The fine tuned model performs best when the prompts contains the key words present in the dataset during fine tuning

Example: Anime Portrait, A boy with colored hair etc

Fine tuned model shows a slight distortion after encountering a new keyword that is not present in the dataset

Example: A boy proposing to a girl

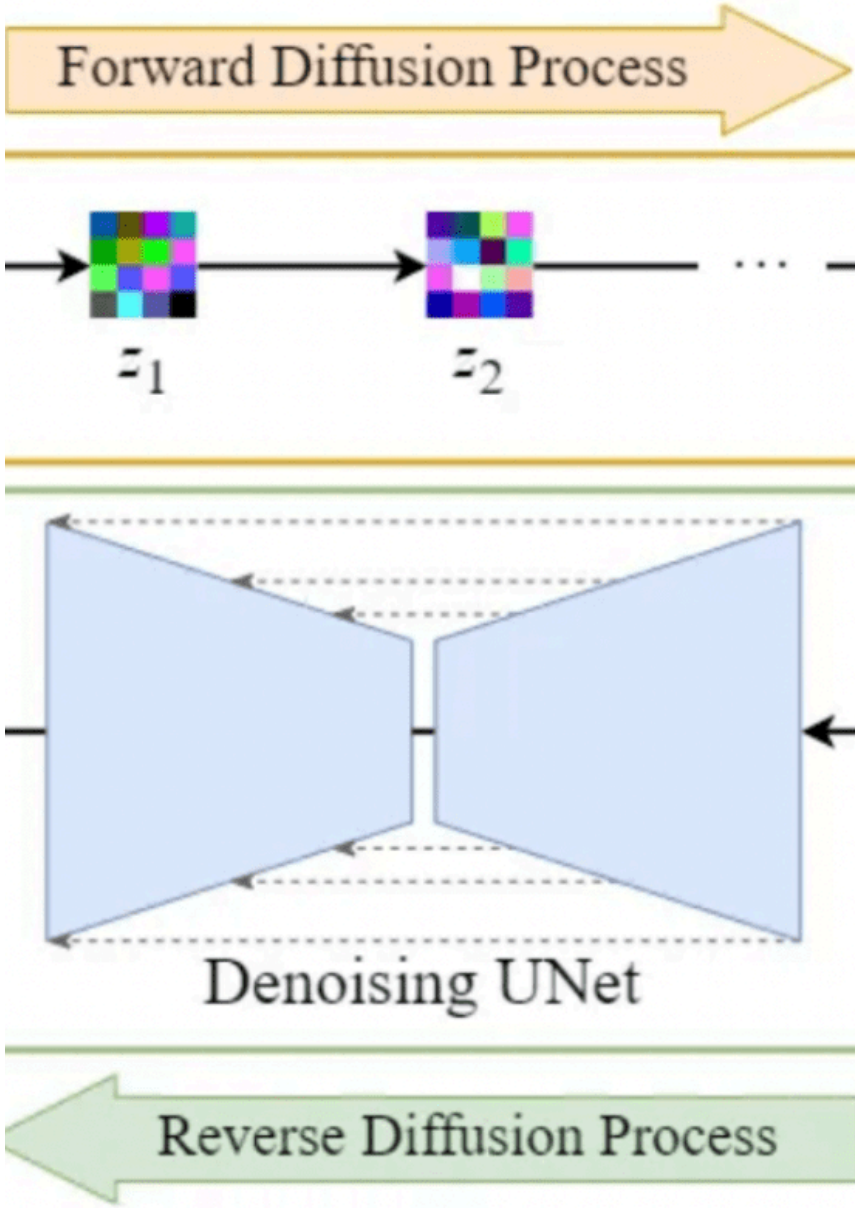


Fine tuned model is not able to fully animate images which are not present in the dataset. Example cat and other animal picturee

Example: A cat wearing glasses

CONCLUSION

The Latent Diffusion Model developed by the team demonstrates the power of combining CLIP tokenization, VAE encoding/decoding, and U-Net architecture to generate high-quality images from text prompts. The model's performance on the Naruto BLIP captions dataset highlights its potential for a wide range of image generation applications. Further research and optimization could lead to even more impressive results, paving the way for more immersive and creative user experiences.



References

1. U-Net: Convolutional Networks for Biomedical Image Segmentation paper
2. https://paperswithcode.com/method/u-net
3. U-Net-and-a-half: Convolutional network for biomedical image segmentation using multiple expert-driven annotations
4. https://arxiv.org/abs/2112.10752

THANK YOU