Trinity International College

(Under the affiliation of Tribhuvan University)

Dillibazar Height, Kathmandu, Nepal

A Project Proposal on

"Text-to-Image Generation"

Submitted to:

Department of Computer Science and Information Technology

Trinity International College

Submitted by:

Neha Shrestha (24287 / 7th Semester / 2076)
Norden Ghising Tamang (24290 / 7th Semester / 2076)

October 16, 2023

Table of Contents

1. INTRODUCTION

Generative AI is such type of artificial intelligence which can be used to generate new content similar to the existing data. The models can create new images, text, audio, video, or other forms of content based on patterns and information present in the data they were trained on. This is one of the popular scopes of machine learning where the machine is trained through different learning approaches to lay foundation models for various unlabeled data.

One of the base examples of the model is the Stable Diffusion. This allows users to generate highly realistic and complex images given a text input. The main idea behind this is to systematically decompose the data structure through an iterative forward diffusion process. Then a reverse diffusion process is applied that restores structure in the data, yielding a highly flexible and tractable model of the data. This diffusion is a latent text-to-image diffusion model that operates by repeatedly reducing noise in a latent representation space and then converting that representation into a complete image.

In this project, we propose a model where thousands of images are registered with layers of noise to form pure noise images. These noisy images are then trained back to get the original images through the use of neural networks. The network for removing the noise is learned by the machine so confidently that the model can be used to turn any random noisy input into a new image which lies in our training data. Through this process, given any text prompt, the model generates the required image.

2. PROBLEM DEFINITION

The problem of data scarcity in data augmentation can be solved through generative AI where huge amounts of training data can be built. Noise injection technique can be used in audio data augmentation improving the model performance. Moreover, visual cues are the base for effective communication along with textual information. The text-to-image model helps to enhance communication and increase learning and understanding skills. Additionally, the developers and graphic professionals need to make a quick visual ideation of the concepts for

clarity and conciseness. The model provides a fast and intuitive way to transform textual ideas into visual prototypes.

3. OBJECTIVE

The objectives of this project are:

- To transform textual descriptions into visually appealing images.
- To ensure the generated images are coherent and contextually accurate.
- To provide a user-friendly interface for users to input text and retrieve generated images.

4. RESEARCH METHODOLOGY

4.1. Literature Review

In the article “Denoising Diffusion Probabilistic Models” the authors took inspiration from nonequilibrium thermodynamics and found a correlation between diffusion probabilistic models and denoising score matching with Langevin dynamics. The parameterization of the diffusion models claims to be the primary contribution to the best sample quality results. The sampling procedure of the diffusion model is discovered to be progressively decoding often resembling autoregressive decoding. [3]

In a study by Esser et al. [2] efficiency of synthesizing high-resolution images is maximized by integrating the usage of transformers as well as CNNs. The power of the inductive bias of CNNs with the articulation of transformers enabled to model and thereby produce quality images. The approach is to use a convolutional VQGAN to learn a codebook of context-rich visual parts, whose composition is subsequently modelled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality.

With diffusion models explained as a sequential technique, the typical operation is performed in pixel space. The formulation algorithm executes by adding or removing noise to a tensor of the same size as the original image resulting in slow inference speed and high computational cost. Thus, the paper “High-Resolution Image Synthesis with Latent Diffusion Models” breaks the ice and deals with issues of previous approaches through the use of latent space of powerful pre-trained autoencoders. The researchers used the models that can be interpreted as an equally weighted sequence of denoising autoencoders which can be trained to predict a denoised variant of their input. The cross-attention conditioning mechanism is used to train a large 1.4 billion parameter text image diffusion model. This model consists of the U-Net and the transformer backbone which are jointly trained on the publicly available LAION 400M dataset. The resulting model is able to compose samples from complex text prompts and also write user-specific text. [1]

4.2. The Framework Of The Model

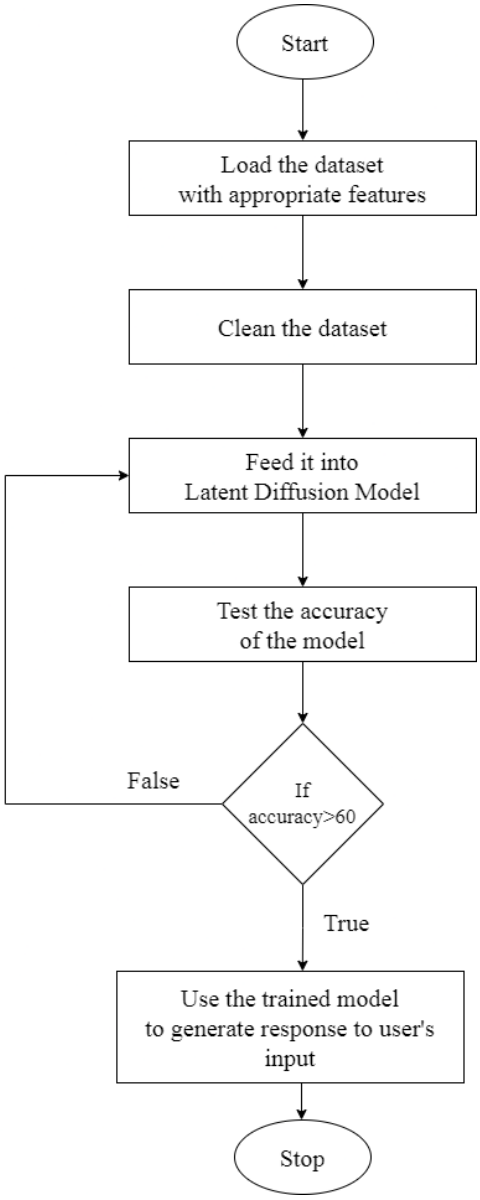

Figure 1: Framework of the model

Figure 1 is the text-to-image generator model's framework. The process initiates by loading the dataset with suitable features, followed by dataset cleaning. After this cleansing step, the data is fed into the Latent Diffusion Model (LDM). The model's effectiveness is evaluated. If the dataset achieves an accuracy surpassing 60%, the trained model is employed to produce an image based on provided textual cues. Conversely, if the accuracy falls short of 60%, the data undergoes further cleaning to attain the accurate response.

4.3. Algorithms / Methods / Theory

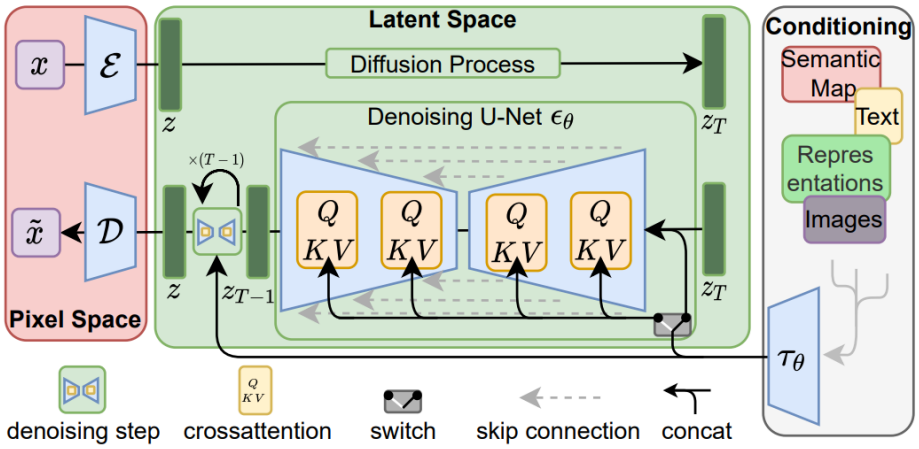

Figure 2: Latent Diffusion Architecture

Figure 2 is the architecture of the Latent Diffusion Model which we are going to implement in our project. The important components of this architecture are listed below:

i. Autoencoder:

An autoencoder is a neural network used for data compression and reconstruction, consisting of an encoder and a decoder. It compresses input into latent space, reconstructs it and aids in training diffusion models on latent space.

ii. Cross-attention Conditioning:

Cross-attention layers turns diffusion models into powerful and flexible generators for general conditioning inputs such as text or bounding boxes and high-resolution synthesis becomes possible in a convolutional manner.

iii. Loss function:

For the loss function, we will be using Mean Squared Error (MSE) loss given by the following formula:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}[\| \epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y)) \|_2^2]$$

iv. Denoising U-Net:

Denoising U-Net is a technique used in LDMs to predict and remove noise from the noisy latent space, allowing for efficient image generation from the latent space with a single network pass, thereby reducing complexity and ensuring perceptually equivalent image space learning.

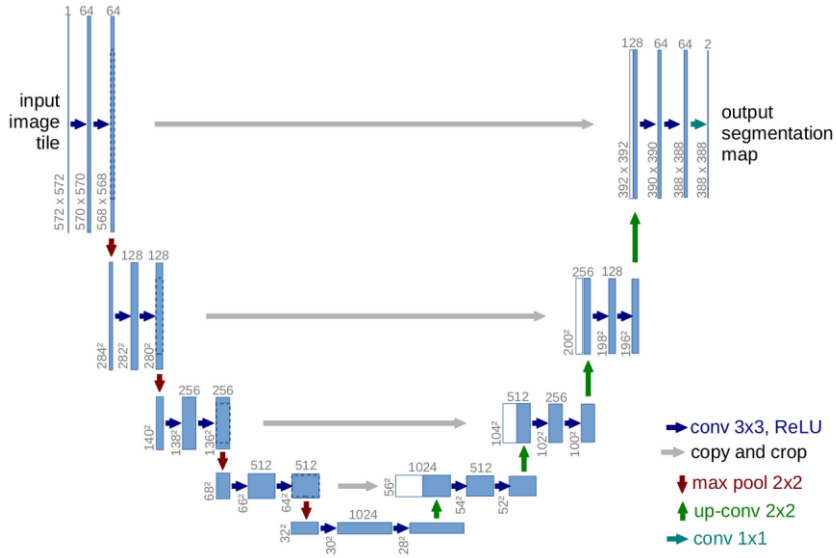

Figure 3: U-Net Architecture

Training Process:

Step-1: Start with an input image.

Step-2: Encode the input image into a compressed latent space using an autoencoder.

Step-3: Add Gaussian noise to the latent space to create a noisy latent space.

Step-4: Use a diffusion model to remove noise from the noisy latent space to retrieve the original latent variables.

Step-5: Use a decoder to reconstruct the image from the original latent variables using text condition which maps into Denoising U-Net.

Step-6: Train the model on a learned latent space to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity.

4.4. Wireframe

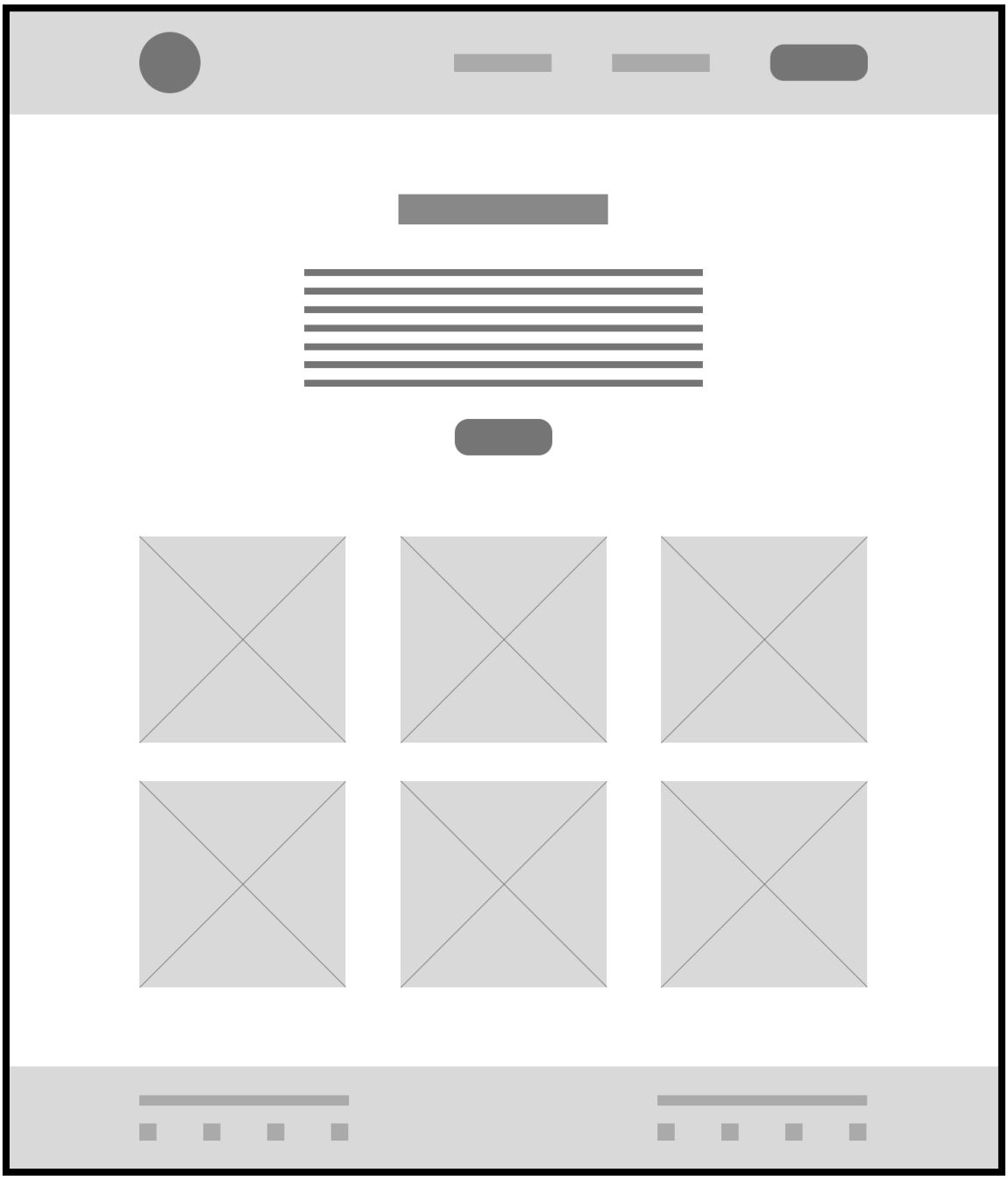

Figure 4(a): Homepage Wireframe of the system

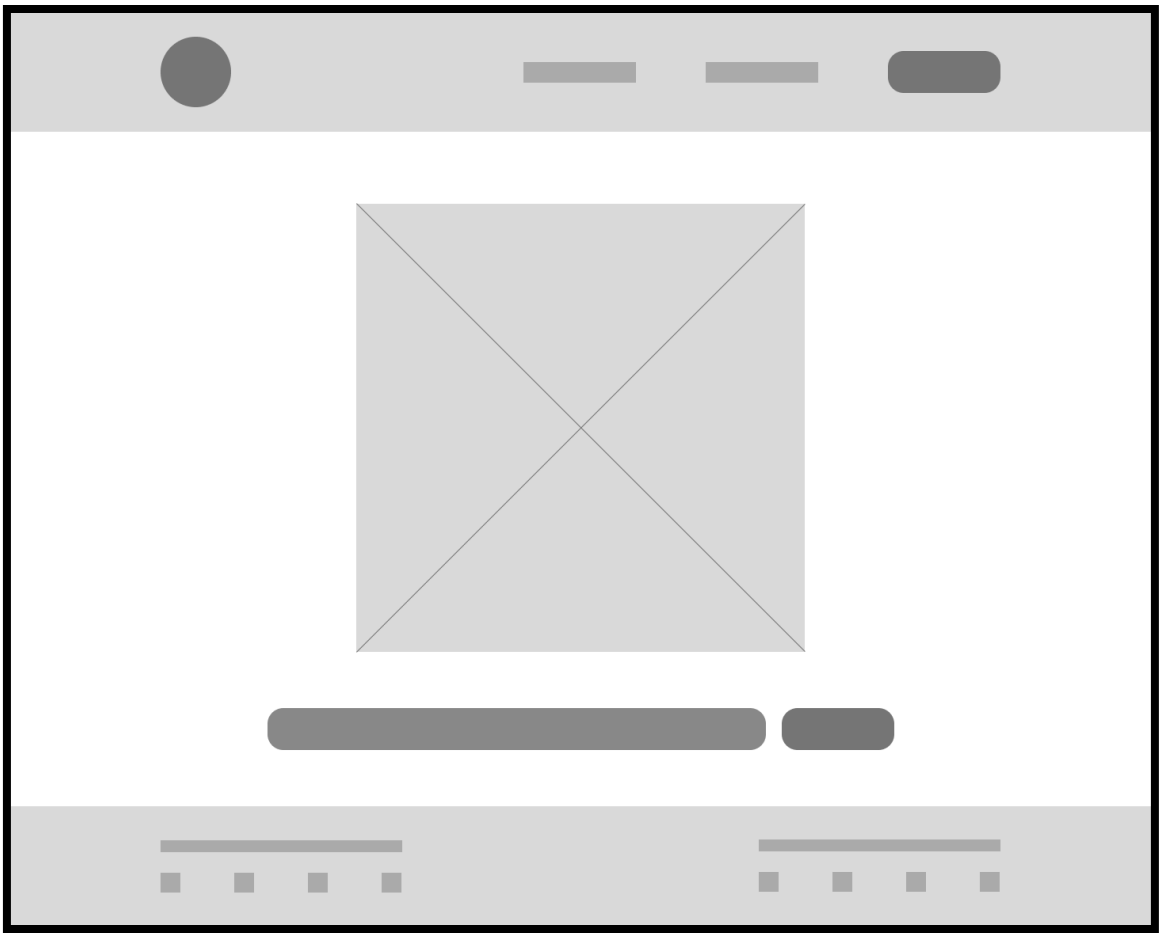

Figure 4(b): Prompting Wireframe of the system

Figure 4(a) and Figure 4(b) are the wireframes of the text-to-image generator model. This is created through Figma and it illustrates the general UI design of our application. The prototype design shows how an end user can interact with the system.

4.5. Mind Map

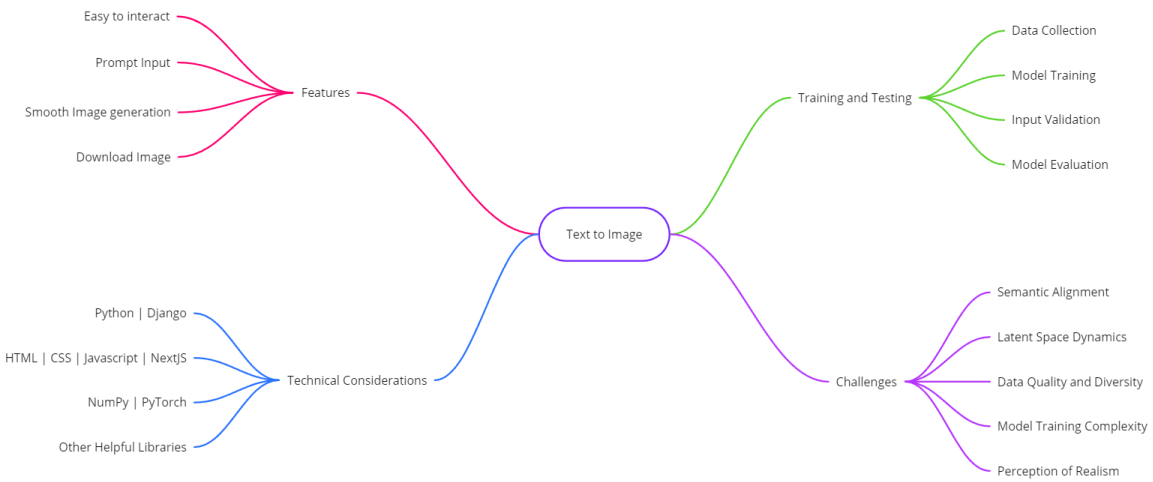

Figure 5: Mind Map of the System

Figure 5 is the Mind Map of the text-to-image generator model. This highlights the major four areas of the model: Features, Technical Considerations, Training and Testing and Challenges. Each area has at least four points to cover through this project.

5. DATA COLLECTION

To train and validate the model, we will collect a dataset of text descriptions associated with corresponding images. Data collection is done by following two methods:

a. Online Dataset:

   Free to use non-commercial datasets such as LAION, ImageNet image-text pairs datasets and various other datasets.

b. Web Scraping:

   Python web scraping scripts will crawl the web and look for image-text pairs.

6. TESTING AND VERIFICATION

Testing involves assessing whether the Text-to-Image generation app performs its intended functions correctly. This can be divided into several components:

i. Input Validation: Verify that the app correctly processes user input, handles potential errors in the textual descriptions.

ii. Image Generation: Confirm that the generated images align with the input text and are contextually relevant. Ensure that the generated images maintain consistency and coherence with the descriptions.

iii. User Interface (UI) Testing: Evaluate the user interface for usability, responsiveness, and intuitive design. Check for any bugs or glitches in the application's graphical user interface.

For Verification, we verify using following measures:

i. Perceptual Metrics: Fréchet Inception Distance (FID) can quantitatively evaluate the visual quality of the generated images.

ii. User Evaluation: Conduct user evaluations where experts or users rate the generated images for aspects such as realism, relevance to the input text, and overall quality. Collect qualitative feedback.

7. EXPECTED OUTPUT

The outcome of this project is to accept user-provided text input, create images based on the input and show the user the generated images.

8. GANTT CHART

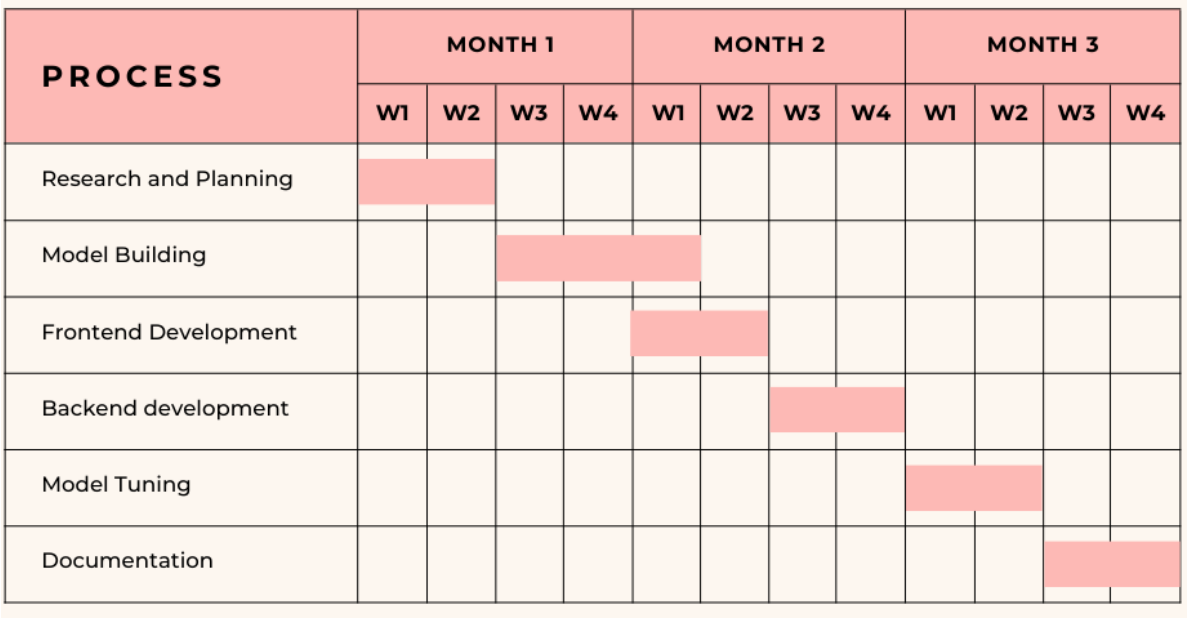

Figure 6: Gantt Chart of the project

Figure 6 shows the Gantt chart of text-to-image generator model. This indicates how the project is manifested to be completed within 3 months. Each activity is allotted a certain duration of time to be fulfilled.

9. REFERENCES

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Bjorn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", *Arxiv*, Apr. 13, 2022. [Online]. Available: https://arxiv.org/pdf/2112.10752.pdf

[2] Patrick Esser, Robin Rombach, Bjorn Ommer, "Taming Transformers for High-Resolution Image Synthesis" *Arxiv*, Dec. 16, 2020. [Online]. Available: https://arxiv.org/pdf/2006.11239.pdf

[3] Jonathan Ho, Ajay Jain and Pieter Abbeel, "Denoising Diffusion Probabilistic Models", *Arxiv*, Dec. 16, 2020. [Online]. Available: https://arxiv.org/pdf/2006.11239.pdf