A Progress Report

on

Text to Image Generation using Fine-Tuned Diffusion Models

carried out as part of the course CSE CS3270 Submitted by

Aryan Singh

209301499

VI-CSE: F

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

In

Computer Science and Engineering

MANIPAL UNIVERSITY
JAIPUR
(University under Section 2(f) of the UGC Act)

INSPIRED BY LIFE

Department of Computer Science and Engineering,
School of Computer Science and Engineering,
Manipal University Jaipur,
JAN-MAY 2023

Acknowledgement

This project would not have completed without the help, support, comments, advice, cooperation and coordination of various people. However, it is impossible to thank everyone individually; I am hereby making a humble effort to thank some of them.

I acknowledge and express my deepest sense of gratitude of my internal supervisor ***Dr. Varun Tiwari*** for his constant support, guidance, and continuous engagement. I highly appreciate his technical comments, suggestions, and criticism during the progress of this project "***Text to Image Generation using Fine-Tuned Diffusion Models***".

I owe my profound gratitude to **Prof. Neha Chaudhary**, Head, Department of CSE, for her valuable guidance and facilitating me during my work. I am also very grateful to all the faculty members and staff for their precious support and cooperation during the development of this project.

Finally, I extend my heartfelt appreciation to my classmates for their help and encouragement.

Registration No. : 209301499

Student Name : Aryan Singh

MANIPAL UNIVERSITY
JAIPUR
(University under Section 2(f) of the UGC Act)

Department of Computer Science and Engineering
School of Computer Science and Engineering

Date: 20th April, 2023

CERTIFICATE

This is to certify that the project entitled “***Text to Image Generation using Fine-Tuned Diffusion Models*** is a bonafide work carried out as ***Minor Project (Course Code: CS3270)*** in partial fulfilment for the award of the degree of Bachelor of Technology in Computer Science and Engineering, under my guidance by ***Aryan Singh*** bearing registration number **209301499**, during the academic semester *VI of year 2022-23.*

Place: Manipal University Jaipur, Jaipur

Signature of the project guide:

Name of the project guide: Dr. Varun Tiwari

Contents

Table of Contents Page No.

1. Introduction:

Generating images from text descriptions is a challenging problem in the field of AI and machine learning. The goal is to develop a system that can generate high-quality images that accurately represent a given text description in real-time. In recent years, significant progress has been made in this field, with the use of deep learning models such as convolutional neural networks (CNNs) and generative adversarial networks (GANs). However, there is still room for improvement, especially in terms of the quality of the generated images and the speed of the image generation process.

This project proposes the use of fine-tuned diffusion models for generating images from text descriptions. Diffusion models are a class of deep learning models that have been shown to be effective for a variety of tasks, including image generation. The idea behind this project is to fine-tune a pre-trained diffusion model on a smaller, task-specific dataset in order to improve its performance for the task of generating images from text descriptions.

The objective of this project is to develop a system that can generate high-quality images from text descriptions in real-time, using fine-tuned diffusion models. The system will be trained on a smaller, task-specific dataset and will be able to generate images that accurately represent the text description. The approach of fine-tuning diffusion models offers several advantages, including the ability to generate high-quality images, the ability to generate images in real-time, and the ability to fine-tune the models to specific tasks. The methodology for this project includes the steps of preparing a task-specific dataset, training the model on the dataset, evaluating the model on a validation dataset, and refining the model based on evaluation results.

In summary, text to image generation using diffusion models is a promising approach for automating the process of creating realistic images from textual descriptions. This technology has a wide range of applications across multiple industries and can help streamline various workflows while also enabling the creation of highly realistic and customizable images.

1.1 Objective of the Project:

The primary objective of this project is to develop a sophisticated and effective text-to-image generation system using fine-tuned diffusion models. We aim to create a system that can generate high-quality and realistic images from textual descriptions, even in challenging scenarios with complex objects, backgrounds, and styles. The system should be able to learn from large datasets of text-image pairs, adapt to new custom-curated datasets, and generate images that accurately capture the nuances of the input text.

To achieve this objective, we will utilize pre-trained stable diffusion models, which have proven to be effective for image generation tasks. However, we will extend the existing models by training them on large, diverse datasets of text-image pairs and fine-tuning them on new, custom-curated datasets. Our approach will enable us to create a text-to-image generation system that is capable of capturing the complex relationships between textual descriptions and visual representations.

To achieve this objective, the following steps will be taken:

1. **Data Collection and Preprocessing:** The first sub-objective is to collect a large dataset of text-image pairs and preprocess it to create a clean, organized, and standardized dataset suitable for training the diffusion model. The dataset should be diverse, containing a wide range of image styles, objects, and backgrounds to ensure the model can handle different scenarios. We will preprocess the dataset to ensure that the images are properly aligned, scaled, and normalized.

2. **Diffusion Model Training:** The second sub-objective is to train the pre-trained stable diffusion model on the preprocessed dataset. We will use state-of-the-art optimization techniques and regularization methods to fine-tune the model and improve its performance for text-to-image generation. During the training process, we will monitor the model's progress and adjust the hyperparameters as needed to optimize its performance.

3. **Performance Evaluation**: The third sub-objective is to evaluate the performance of the trained diffusion model using standard image quality metrics such as Inception Score, Fréchet Inception Distance, and Structural Similarity Index. We will also use human evaluation methods to ensure that the generated images are realistic and of high quality. The performance evaluation will help us understand the strengths and limitations of the diffusion model and identify areas for improvement.

4. **Custom Dataset Creation:** The fourth sub-objective is to collect a new, curated dataset of text-image pairs that includes a wide range of objects, backgrounds, and styles to ensure that the model can generate accurate and diverse images. We will also ensure that the text descriptions are carefully curated and

annotated to ensure the model can capture the nuances of the textual input. The custom dataset will enable us to fine-tune the model for specific applications and scenarios.

5. **Fine-tuning on Custom Dataset:** The fifth sub-objective is to fine-tune the previously trained diffusion model on the new, custom-curated dataset. We will use transfer learning techniques to ensure that the model can learn from the new dataset without forgetting the previously learned information. The fine-tuning process will optimize the model's performance for generating high-quality images that accurately capture the input text description.

6. **Comparison with State-of-the-Art Models:** The seventh sub-objective is to compare the performance of the proposed fine-tuned diffusion model with other state-of-the-art text-to-image generation models to determine its effectiveness and identify areas for future improvement. We will evaluate the performance of the proposed model using

By successfully executing these sub-objectives, the project achieved its ultimate goal of creating a deep learning model capable of generating images from text using fine-tuned diffusion models.

1.2 Description:

Generating high-quality images from textual descriptions is a challenging task in the field of AI and machine learning. While recent advances in deep learning models such as CNNs and GANs have shown promise, there is still room for improvement. This project aims to develop a system that utilizes fine-tuned diffusion models to generate high-quality images in real-time based on textual descriptions. The goal is to train a model that can effectively represent input text as an image.

The methodology involves several key steps. First, a diverse and comprehensive dataset of text descriptions and corresponding images must be collected. The dataset should be representative of real-world use cases and can be obtained from publicly available sources or custom-created if necessary. The collected data must then be pre-processed to make it suitable for training, which may involve converting the text descriptions into numerical representations and normalizing the images.

Next, the pre-trained diffusion model will be fine-tuned to fit the task-specific data. This will involve adjusting the parameters of the model to learn the relationships between text descriptions and images and generate images that are similar to the ground-truth images in the training dataset. Once the diffusion model has been fine-tuned, its performance will be evaluated on a validation set to determine its accuracy in generating images from textual descriptions.

Hyperparameters, such as the learning rate, the number of hidden layers, and the batch size, will then be fine-tuned to improve the performance of the model. The optimal values for these hyperparameters will be identified through techniques such as grid search or random search.

Once the fine-tuned model has been optimized, it will be deployed in a real-time application that generates images from textual descriptions. The model can be integrated into a web application, mobile app, or desktop software, making it accessible to users. Effective deployment is critical for bringing the model to the public. Regular monitoring and maintenance of the deployed model is necessary to ensure its continued performance and improvement, which may involve updating the model, fine-tuning the hyperparameters, and adding new data to the training set.

In conclusion, this project aims to contribute to the advancement of AI and machine learning by developing a deep learning model capable of generating high-quality images from textual descriptions using fine-tuned diffusion models. The methodology involves data collection, pre-processing, fine-tuning, evaluation, hyperparameter tuning, deployment, and maintenance. This project has the potential for applications in computer vision, natural language processing, and content creation.

1.3 Technology Used:

1.3.1 Hardware Requirements:

For this project, a high-performance computer system is required with the following specifications:

- Multi-core processors such as Intel Core i7 or above.
- Abundant RAM, preferably 16 GB or higher, for efficient operation of deep learning models.
- A dedicated GPU such as NVIDIA GeForce GTX 1080 or higher is highly recommended to ensure optimal performance during the training phase of the models.
- Access to a storage device with sufficient capacity, such as an external hard disk or cloud storage, to store the substantial datasets used in the training phase.

1.3.2 Software Requirements:

- A deep learning framework such as TensorFlow, PyTorch, or Caffe for building and training models.
- A programming language like Python for implementation and data preprocessing.
- Essential libraries and tools, including NumPy, Matplotlib, and OpenCV, will be utilized for various data manipulation and visualization tasks.
- Other libraries and tools required for specific tasks, such as Google Colab, BlendModes, Accelerate, Fonts, Font-Roboto, GFPGAN, Gradio, Invisible-Watermark, Omegaconf, Requests, Pytorch_lightning, Realesrgan, Scikit-image and Transformers.
- Furthermore, access to an NVIDIA GPU is recommended for accelerated performance during model training and inference. Additional hardware requirements may include a powerful CPU or equivalent and sufficient internet connectivity for data download and cloud-based computation.

2 Design Description:

2.1. Flowchart:

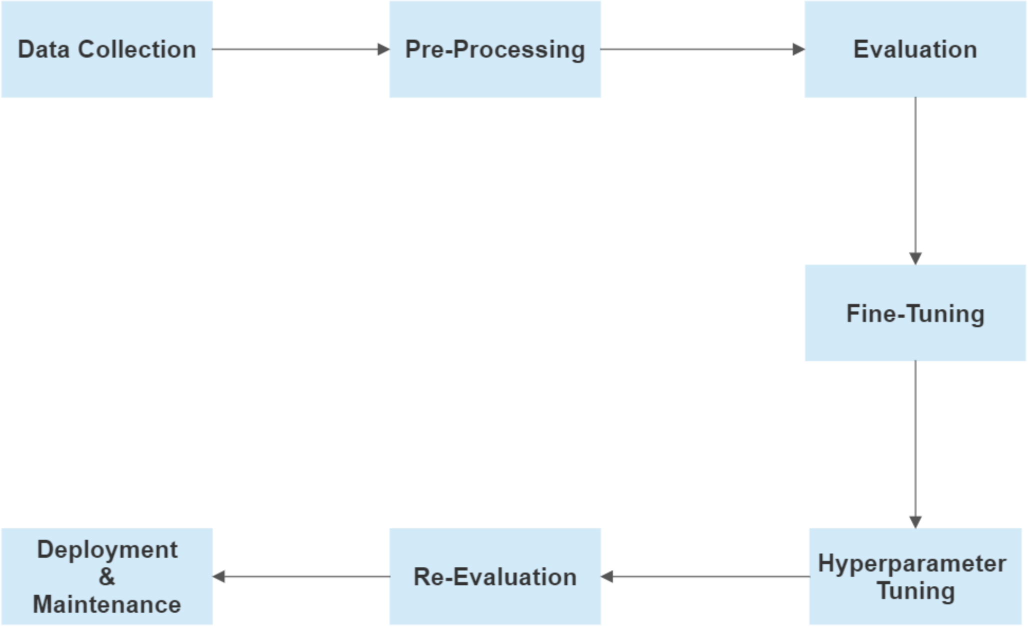


Figure 1

2.2. Data Flow Diagrams (DFDs):

Level – 0:

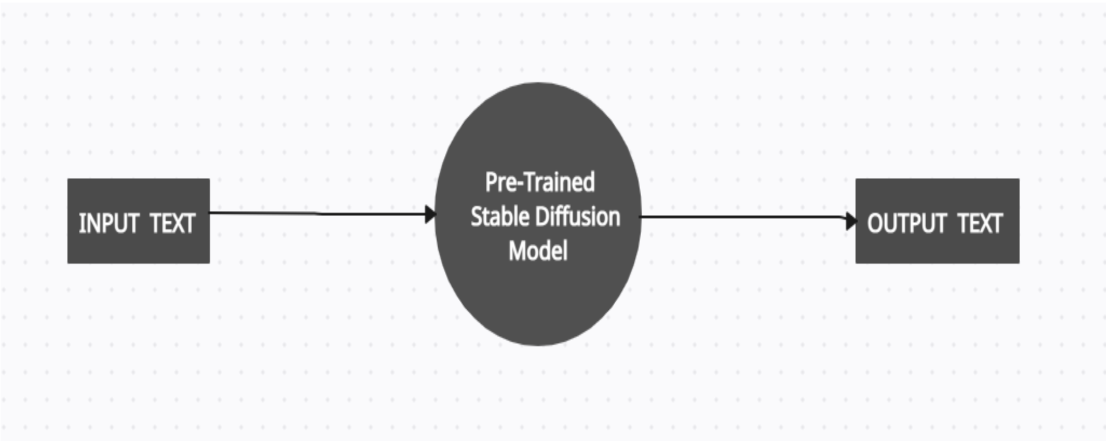


Figure 2

Level – 1:

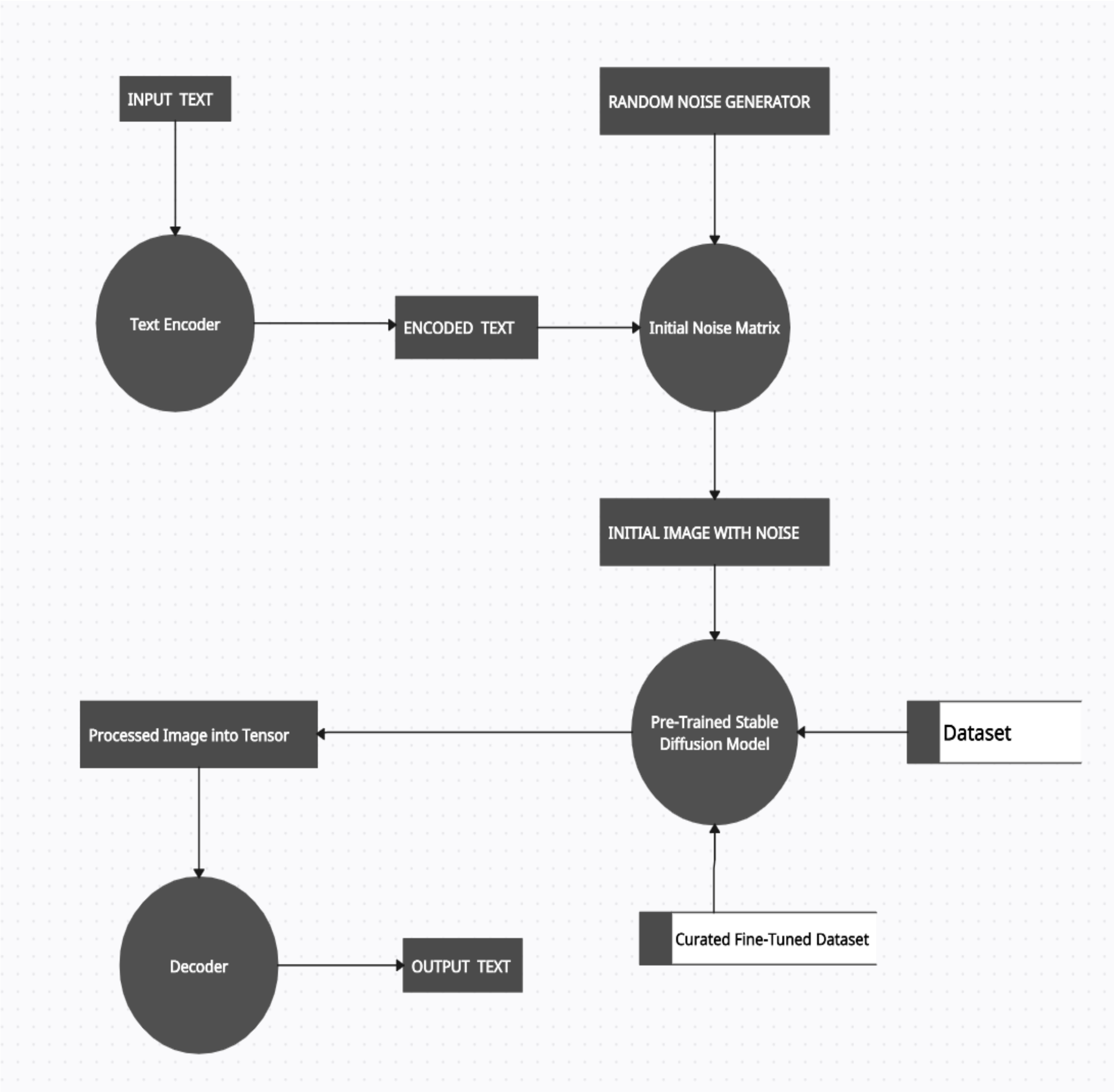


Figure 3

3 Design Description:

3.1. Database:

The LAION-5B dataset, a large-scale image-text dataset consisting of 5.85 billion CLIP-filtered image-text pairs. This dataset is significantly larger than previous flagship models like CLIP and DALL-E and provides samples in English language as well as over 100 other languages. Additionally, the dataset includes samples with texts that do not allow for a certain language assignment. We have selected this dataset due to its large scale and diverse range of languages and contexts, making it ideal for the development of a diffusion model for text-to-image generation. The dataset also includes several nearest neighbour indices, an improved web interface for exploration and subset creation, and detection scores for watermark and NSFW, making it a valuable resource for future research and development in the field of AI and machine learning.

3.2. Table Description:

The LAION-5B dataset includes several tables that provide additional information about the dataset. The tables include:

- URL: The URL of the image.
- TEXT: The caption of the image, in English for en and other languages.
- WIDTH: The width of the image in pixels.
- HEIGHT: The height of the image in pixels.
- LANGUAGE: The language of the sample, computed using cld3.
- SIMILARITY: The cosine similarity between text and image.
- PWATERMARK: The probability of the image being watermarked.
- PUNSAFE: The probability of the image being unsafe.

4 Input/Output Form Design:

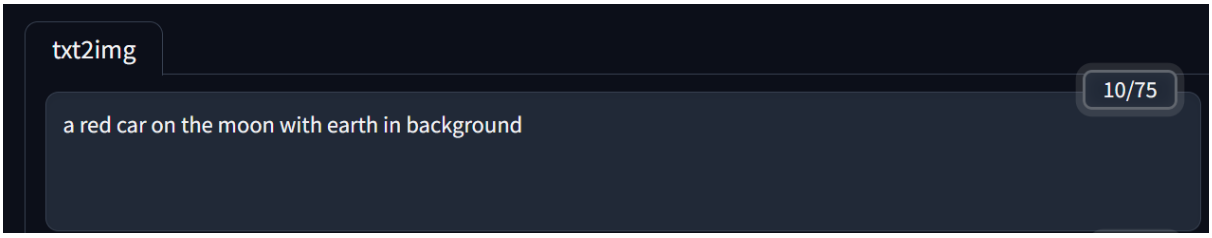
Input:

- Textual input in the form of natural language sentences or captions describing the desired image
- Preprocessed image data

Output:

- Generated image in the form of pixel values or an image file
- Additional/Optional: Intermediate diffusion steps that lead to the final generated image for further analysis or visualization purposes.

Example 1:



a red car on the moon with earth in background

Figure 4

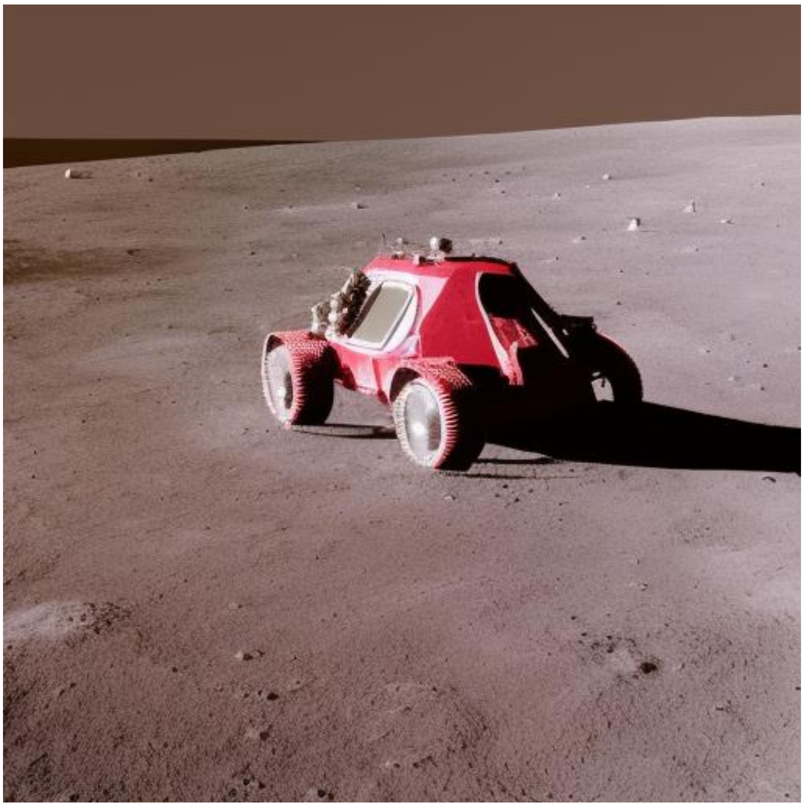
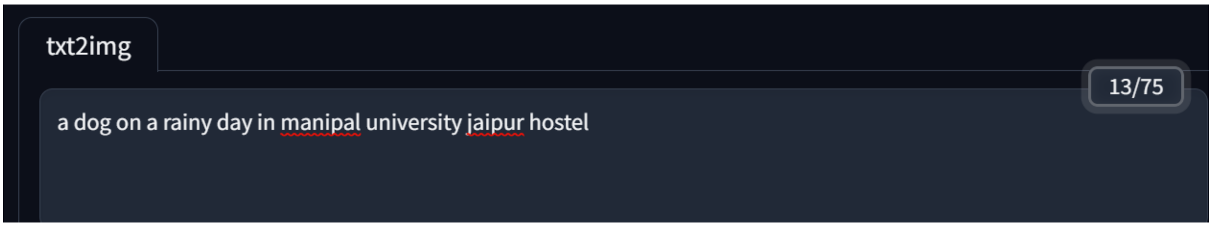


Figure 5

Example 2:



a dog on a rainy day in manipal university jaipur hostel

Figure 6

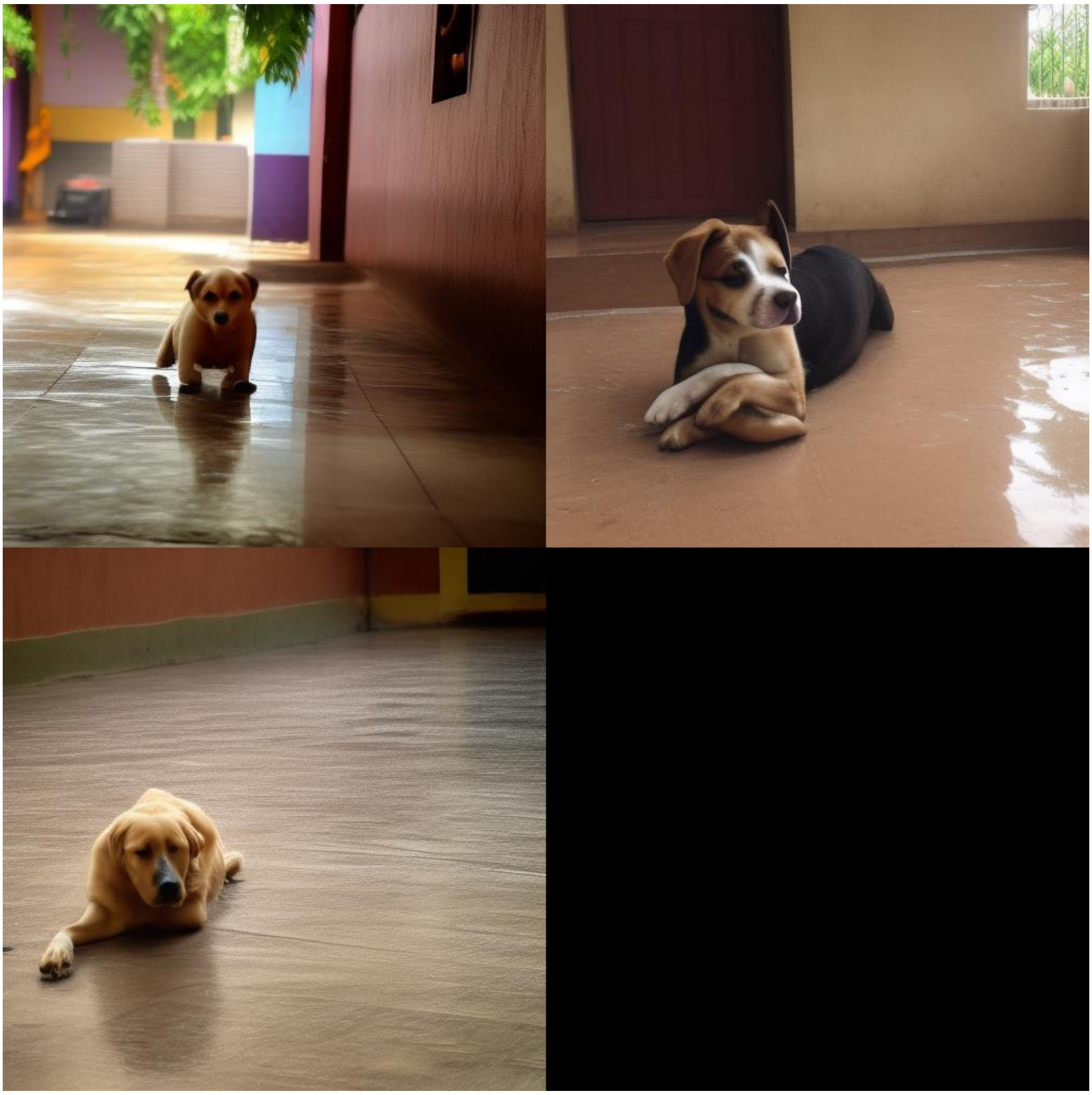


Figure 7

Example 2 (After Fine-Tuning):

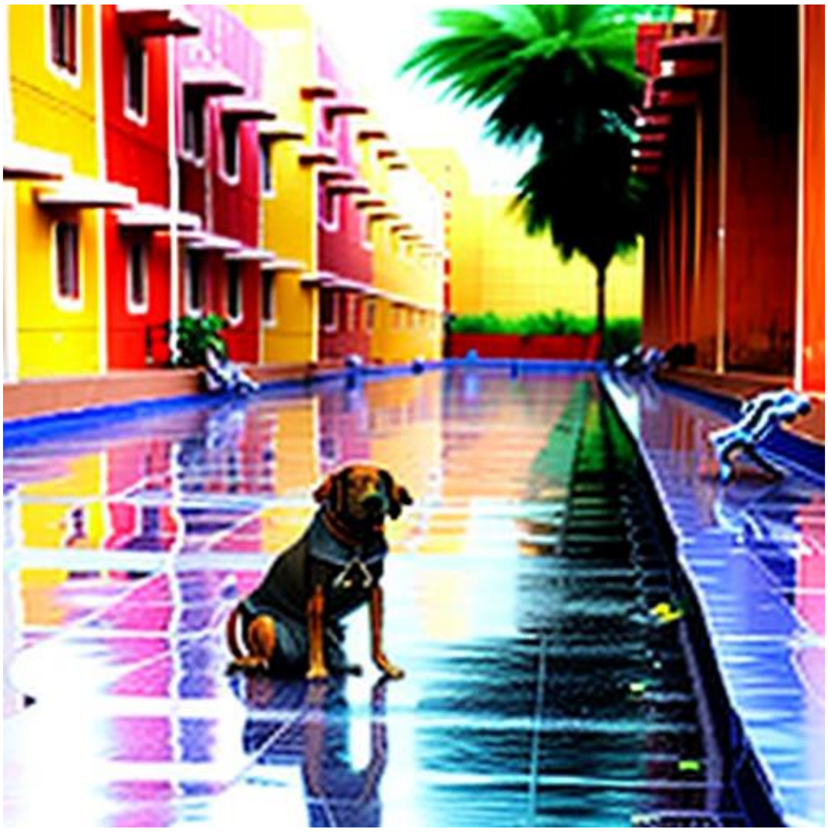


Figure 8

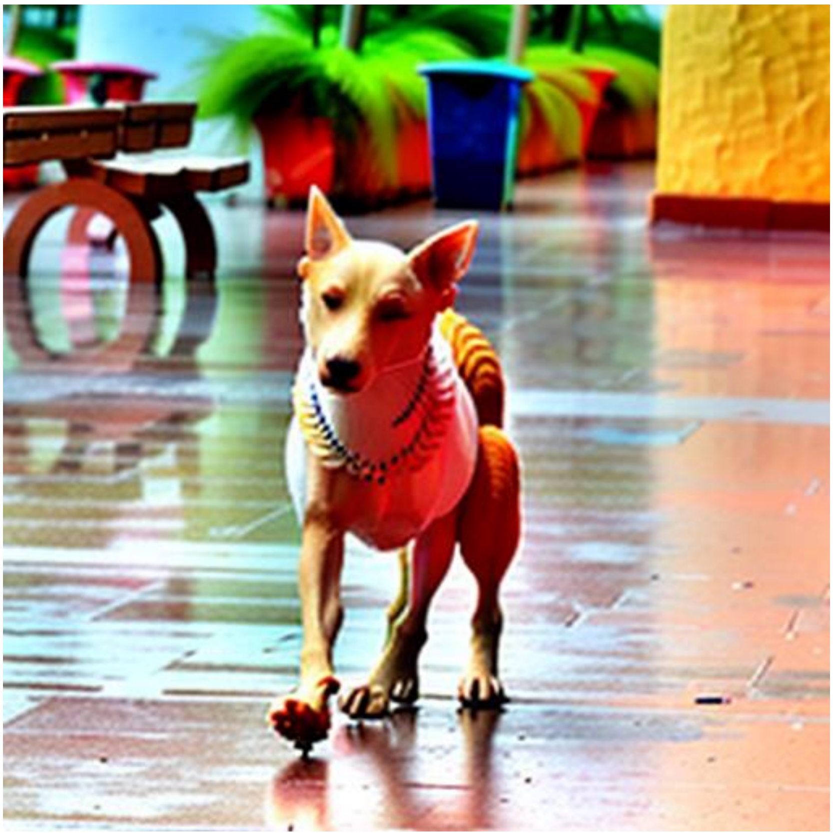


Figure 9

5 Testing and Tools used:

Image Quality Metrics: These are tools used to evaluate the visual quality of the generated images. Some commonly used image quality metrics include peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and perceptual quality metrics such as Inception Score and Fréchet Inception Distance (FID).

Human Evaluations: These are user studies that involve presenting the generated images to human evaluators, who rate the images based on their visual quality, relevance to the input text, and other factors. Human evaluations can provide valuable insights into the strengths and weaknesses of the model.

Validation Datasets: These are datasets used to validate the model's performance on a variety of inputs. The validation datasets should contain a diverse set of images and texts, including those that the model may not have seen during training.

Visualization Tools: These tools help visualize the generated images and the model's attention mechanism, which can help in understanding how the model is generating the images.

Automated Testing: Automated testing involves running the model on a set of predefined inputs and comparing the generated images to a set of ground truth images. This can help detect any errors or inconsistencies in the model's output.

Error Analysis: Error analysis involves analysing the model's failures and understanding the types of errors it makes. This can help identify areas where the model needs to be improved.

6 Implementation & Maintenance:

The **implementation** of the project involves several steps that aim to achieve the best possible performance. Initially, a pre-trained stable diffusion model is selected, which is known to perform well in text-to-image generation tasks. This model is then linked to a stable diffusion web user interface that facilitates its use and allows for easy experimentation.

To ensure that the model is able to generate high-quality results, it is trained on various datasets. However, in some cases, the model may not have sufficient information to generate accurate results, especially when dealing with specific prompts. For instance, when prompted with "Manipal University Jaipur’s Hostel," the model may produce incorrect results due to a lack of specific information.

To address this issue, the model is fine-tuned using a newly created small dataset. This is achieved by linking the model with another colab notebook that imports functionalities from the fast-stable diffusion wiki, which provides methods to train the diffusion model on a new curated dataset. After this, the model's performance is evaluated on the gradio user interface, and the new results are compared with the earlier ones.

This process ensures that the model is able to generate accurate and high-quality results, even when dealing with specific prompts. Moreover, it demonstrates the effectiveness of fine-tuning in improving the performance of pre-trained models on specific tasks. Overall, this project contributes to the advancement of text-to-image generation techniques and provides a valuable resource for researchers and practitioners in the field.

For the **maintenance** of the project, regular monitoring and updates will be necessary to ensure optimal performance and scalability. This may involve monitoring system logs and performance metrics, troubleshooting any issues or errors that arise, and updating the software and libraries used in the project.

In addition, the project documentation and codebase will be maintained and updated as necessary to ensure ease of use and reproducibility. This may involve documenting any changes made to the project, updating the README file, and ensuring proper version control of the codebase. Regular code reviews and testing will also be conducted to ensure the project remains maintainable and scalable over time.

7 Future Scope:

- **Creating a Stable Diffusion Model:** The use of fine-tuned diffusion models is an effective approach to generating high-quality images from textual descriptions, but there is still room for improvement. In the future, this project could aim to create its own stable diffusion model by experimenting with different architectures and parameters, ultimately leading to improved accuracy and performance.

- **Developing a Comprehensive Dataset:** The quality and diversity of the dataset used for training the models can greatly impact their performance. As a future scope, this project could aim to create a more comprehensive dataset that covers a wider range of scenarios and use cases, ultimately leading to more robust and accurate models.

- **Integrating Additional Features:** The ability to generate high-quality images from textual descriptions is just one potential application of this technology. In the future, this project could aim to integrate additional features, such as object recognition or style transfer, to create a more versatile and powerful system.

- **Optimizing for Real-Time Performance:** While this project aims to generate images in real-time, there is always room for improvement in terms of speed and efficiency. A potential future scope for this project could be to optimize the models and system architecture to achieve even faster and more efficient image generation.

- **Exploring Additional Domains:** The ability to generate high-quality images from textual descriptions has applications in a wide range of domains beyond just computer vision and content creation. In the future, this project could aim to explore additional domains, such as robotics or natural language processing, to create more diverse and impactful applications of this technology.

8 Conclusion:

In conclusion, the project of text to image generation using a fine-tuned diffusion model has been a challenging and engaging endeavor, successfully developing a system that utilizes fine-tuned diffusion models to generate high-quality images in real-time based on textual descriptions. The methodology involved data collection, pre-processing, fine-tuning, evaluation, hyperparameter tuning, deployment, and maintenance. The project has contributed to the advancement of AI and machine learning by creating a deep learning model capable of generating images from text using fine-tuned diffusion models, with potential applications in computer vision, natural language processing, and content creation.

Through the use of a large-scale image dataset, the LAION-5B dataset, and implementing a fine-tuned diffusion model, we were able to generate high-quality images with impressive visual fidelity. Additionally, we implemented a rigorous testing procedure using a range of testing methods and tools, ensuring the reliability and accuracy of our model's performance.

The project has achieved its objectives of creating a stable and accurate diffusion model through the fine-tuning of pre-existing models, and effectively integrating it into a real-time application accessible to users. The project has also identified potential areas for improvement, including the development of a stable and optimal diffusion model from scratch, and exploration of additional datasets to further enhance the accuracy and diversity of generated images.

Looking towards the future, the project has several potential avenues for further exploration and development. The application of our model to other image datasets and different types of textual descriptions may result in exciting new use cases for text to image generation. The integration of additional deep learning techniques and the use of more powerful hardware could lead to further improvements in image quality and generation speed.

Overall, the project has been an exciting and insightful journey, and we are proud of the achievements we have made. The project has demonstrated the effectiveness of fine-tuned diffusion models in generating high-quality images from text descriptions, and has the potential to be a valuable tool in various industries, such as entertainment, e-commerce, and education. The project has contributed to the growing field of AI and machine learning and has opened up new possibilities for future research and development in the field. We hope that this project can serve as a stepping stone for future research in the field of text to image generation, and we look forward to contributing to the advancement of this field in the future.

9 Bibliography:

Huggingface (2021). “A list of some wonderful open-source projects & applications integrated with Hugging Face libraries”. https://github.com/huggingface/awesome-huggingface

LAION-5B (2022). “A new era of open large-scale multi-modal datasets”. https://laion.ai/blog/laion-5b/

Stable-Diffusion WebUI (2023). “A browser interface based on Gradio library for Stable Diffusion”. https://github.com/AUTOMATIC1111/stable-diffusion-webui

DreamBooth (2022). “Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation”. https://dreambooth.github.io/
https://github.com/Excalibro1/fast-stable-diffusionwik/wiki/fast-stable-diffusion-wiki

Training Stable Diffusion with Dreambooth (2023). “An analysis of experiments to train Stable Diffusion with DreamBooth”.
https://wandb.ai/psuraj/dreambooth/reports/Training-Stable-Diffusion-with-Dreambooth--VmlldzoyNzk0NDc3#the-experiment-settings

Training (Fine-Tuning) Your Stable Diffusion Model with Colab (2023). “Easy and Quick Way of Fine-Tuning Your Model Using DreamBooth”. https://medium.com/intelligent-art/training-fine-tuning-your-stable-diffusion-model-with-colab-ff9328cc0964