

Abstract geometric lines in black on a white background, forming various overlapping polygons and shapes, primarily concentrated on the left side of the slide.

TEXT TO IMAGE GENERATION USING FINE-TUNED DIFFUSION MODELS

Aryan Singh

(209301499)

AGENDA

Introduction

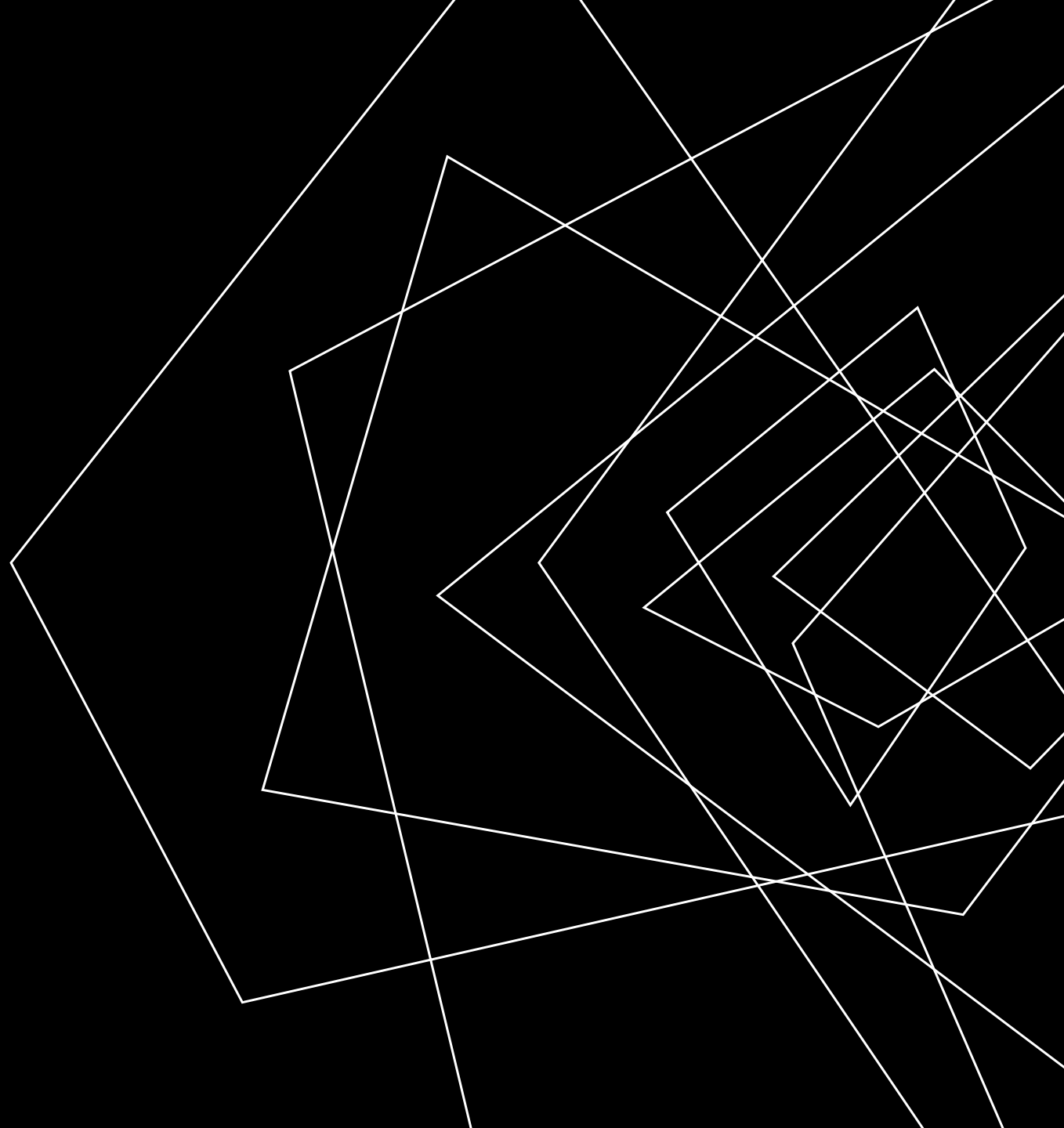
Project Background

Objectives

Dataset

Future Prospects

Conclusion





INTRODUCTION

- Generating high-quality images from text descriptions is a challenging problem in AI and machine learning.
- Deep learning models like CNNs and GANs have shown significant progress in this field, but there is still room for improvement.
- 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 effective for image generation tasks.
- The project aims to fine-tune a pre-trained diffusion model on a smaller, task-specific dataset to improve performance.
- The objective is to develop a system that can generate high-quality images in real-time from text descriptions using fine-tuned diffusion models.

- Fine-tuning diffusion models offer several advantages, including high-quality image generation, real-time generation, and the ability to fine-tune models for specific tasks.
- The methodology involves preparing a pre-trained stable diffusion model, training on a larger dataset, fine-tuning the model and evaluating the result.
- The system will accurately represent the text description in generated images.
- The project addresses the limitations of current methods and aims to improve the quality of generated images and the speed of the image generation process.



PROJECT BACKGROUND

PROJECT BACKGROUND

The field of AI and machine learning has been making remarkable progress in recent years, especially in the domain of image generation. One of the most challenging problems in this domain is generating high-quality images from textual descriptions. To address this problem, several approaches have been proposed, including deep learning models such as convolutional neural networks (CNNs) and generative adversarial networks (GANs). However, there is still room for improvement 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, aiming to achieve higher-quality image generation in real-time.



OBJECTIVES

The aim of this project is to create a system that utilizes fine-tuned diffusion models to generate high-quality images in real-time based on text descriptions. The model will be trained on a curated and task-specific dataset to produce images that effectively represent the input text.

OBJECTIVES

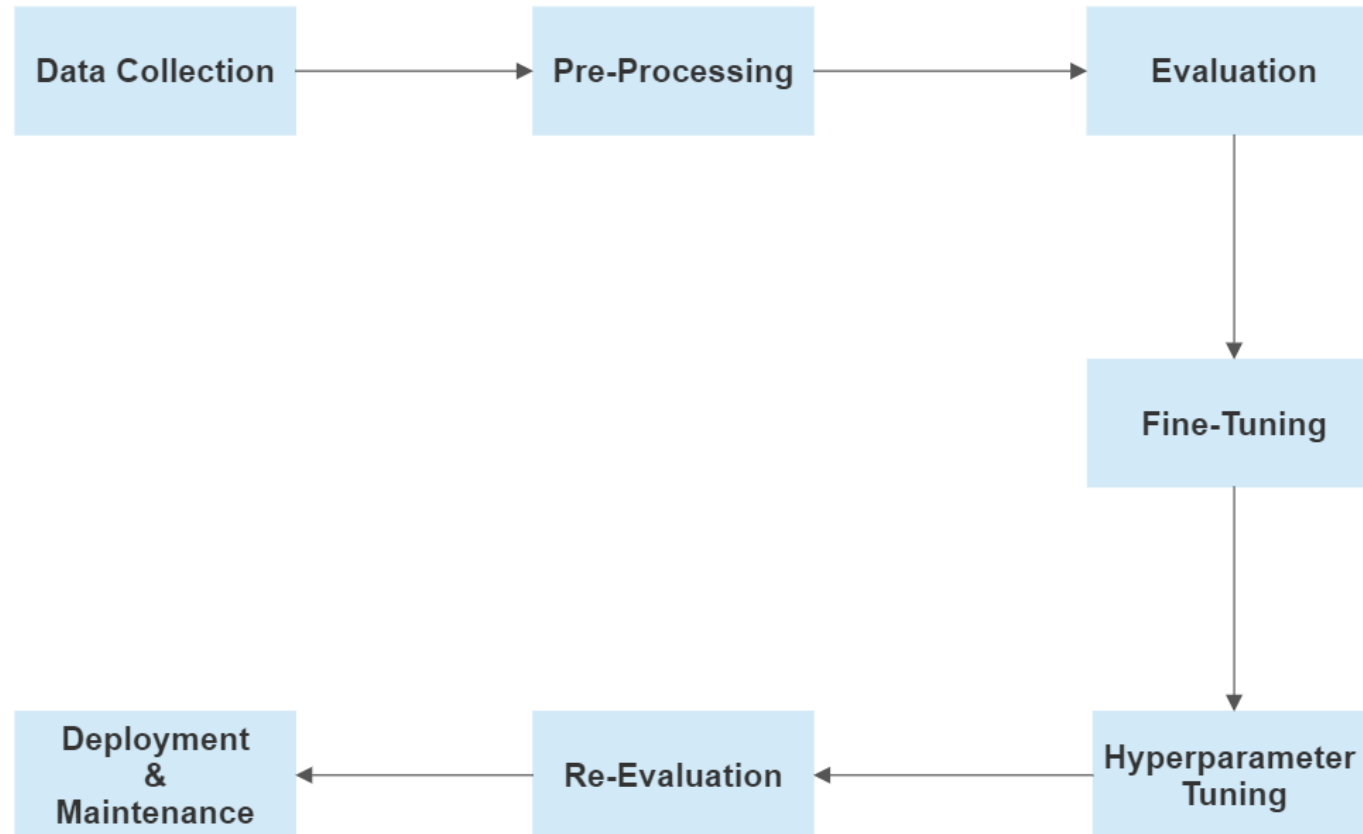
By successfully executing the following sub-objectives, the project will achieve its ultimate goal of creating a deep learning model capable of generating images from text using fine-tuned diffusion models.

- **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.
- **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.
- **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.

OBJECTIVES

- **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.
- **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.
- **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

DEVELOPMENT CYCLE





DATABASE

A series of thin, black, intersecting lines in the top-left corner of the page, creating a geometric pattern.

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 neighbor 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.

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.

DATASET DESCRIPTION

LANGUAGE _____ The language of the sample.

SIMILARITY _____ The cosine similarity between text and image.

PWATERMARK _____ The probability of the image being watermarked.

PUNSAFE _____ The probability of the image being unsafe.

DATASET DESCRIPTION



USER INTERFACE

Stable Diffusion checkpoint
aryan_test_mk3.ckpt [540e663c24]

SD VAE
Automatic

Clip skip
1

Inpainting conditioning mask strength
1

Noise multiplier for img2img
1

txt2img

a dog in manipal university jaipur hostel on a rainy day

13/75

Generate

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

Styles

Sampling method
Euler a

Sampling steps
20

☐ Restore faces

☐ Tiling

☐ Hires. fix

Width
512

Height
512

CFG Scale
13.5

Batch count
6


Batch size
1








Seed
-1

Extra

ControlNet

Script
None





Save

Zip

Send to img2img

Send to inpaint

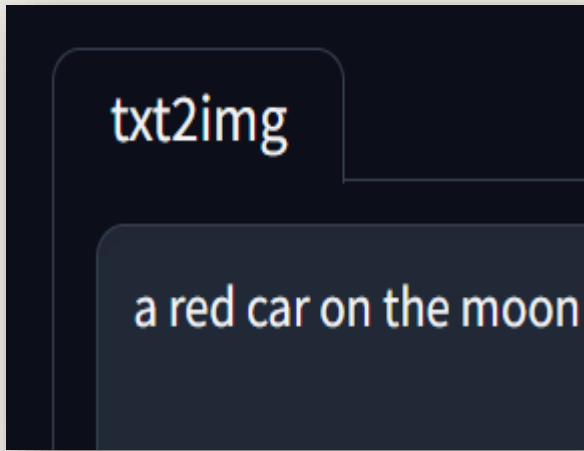
Send to extras

a dog in manipal university jaipur hostel on a rainy day
Steps: 20, Sampler: Euler a, CFG scale: 13.5, Seed: 3056755576, Size: 512x512, Model hash: 540e663c24, Model: aryan_test_mk3



**RESULTS
BEFORE
FINE-TUNING**

INPUT 1



OUTPUT 1



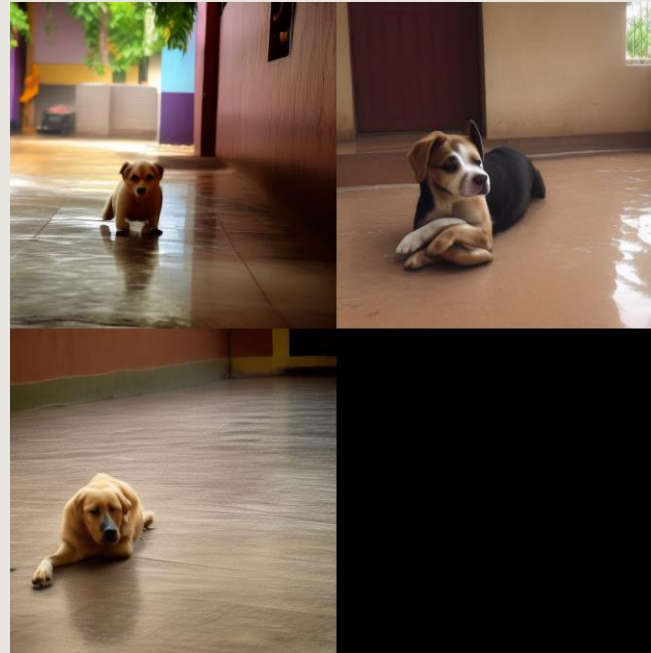
EXAMPLE 1

txt2img

a dog on a rainy day in manipal university jaipur hostel

INPUT 1

OUTPUT 1



EXAMPLE 2



**RESULTS
AFTER
FINE-TUNING**

EXAMPLE 1

INPUT 1

txt2img

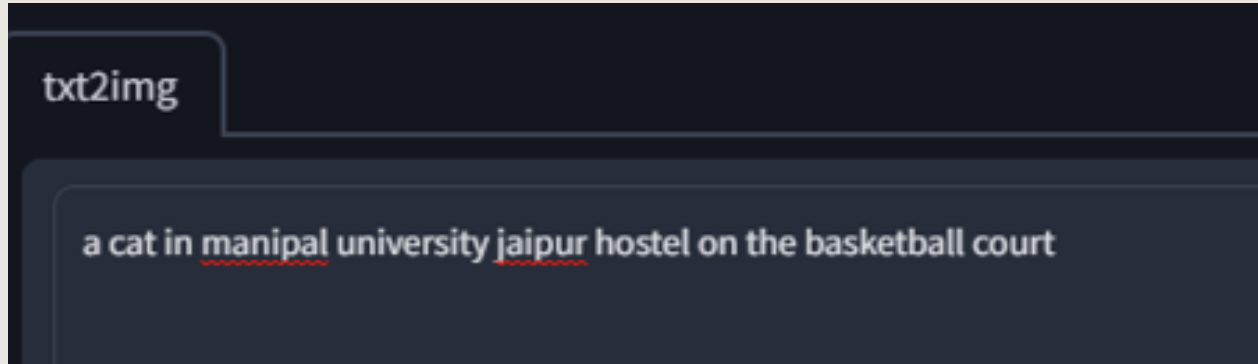
a dog in manipal university jaipur hostel on a rainy day

OUTPUT 1



EXAMPLE 1

INPUT 1




OUTPUT 1

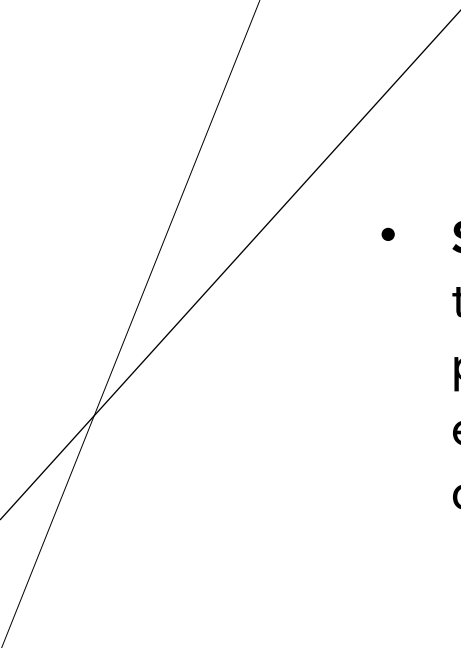


EXAMPLE 2



FUTURE PROSPECTS

- 
- **Improving the quality of generated images:** While the fine-tuned diffusion model may be effective at generating plausible images, there is still room for improvement in terms of image quality. Future research could focus on developing more sophisticated models that can generate more realistic and high-quality images.
 - **Enhancing the model's ability to understand context:** While the model may be able to generate images based on textual descriptions, it may struggle with more nuanced or complex descriptions that require a deeper understanding of context. Future research could explore ways to enhance the model's ability to understand context, such as incorporating additional contextual information or leveraging other natural language processing techniques.
 - **Adapting the model for other applications:** While the project focuses on generating images from textual descriptions, the underlying technique of fine-tuning a diffusion model could potentially be applied to other domains as well. Future research could explore how the same approach could be used for tasks like video generation, music generation, or even natural language generation.

- 
- **Scaling the model to handle larger datasets:** While the COCO dataset used in the project is quite large, there are even larger datasets available that could potentially be used to train a more powerful model. Future research could explore ways to scale the model to handle these larger datasets and take advantage of the additional training data.
 - **Integrating the model into real-world applications:** While the project focuses on generating images in a research context, there are many potential real-world applications for a text-to-image generation model. Future research could explore ways to integrate the model into applications like e-commerce, where it could be used to generate product images based on textual descriptions, or in virtual reality or gaming, where it could be used to generate realistic 3D environments based on textual descriptions

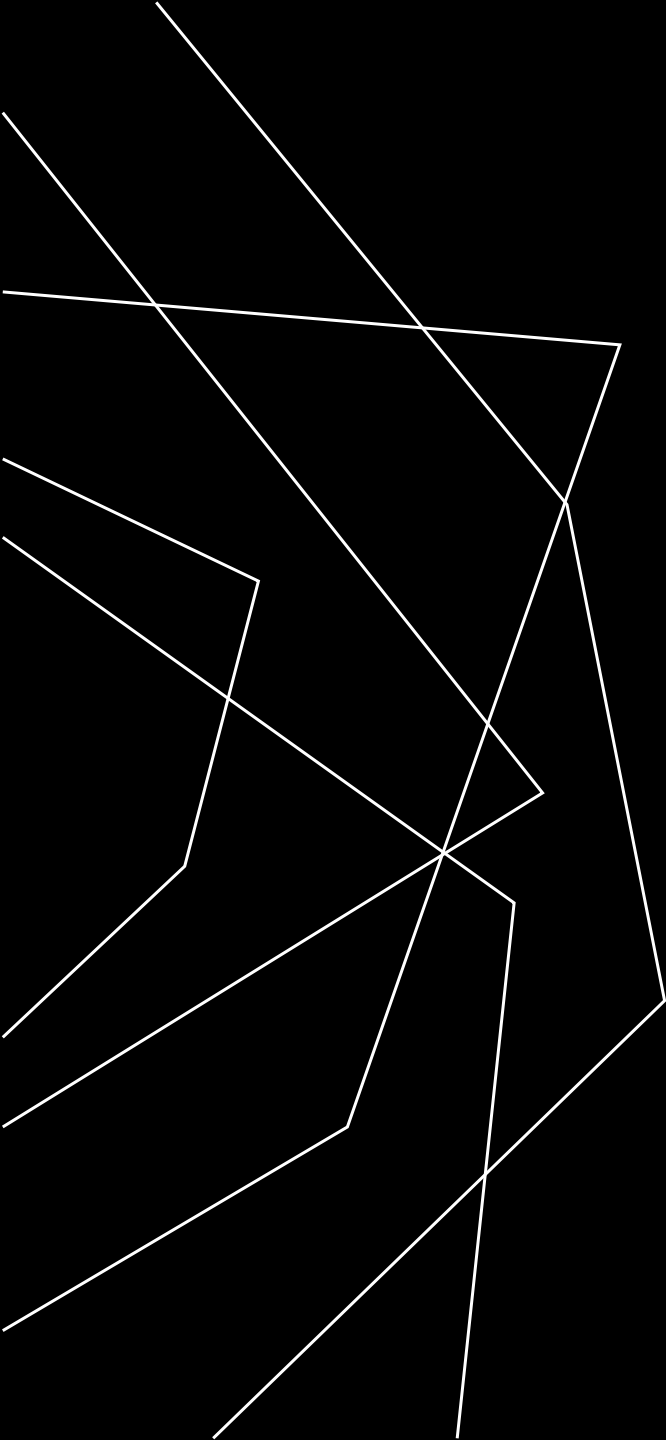


CONCLUSION



CONCLUSION

The use of fine-tuned diffusion models for image generation from text descriptions has shown great potential for achieving high-quality image generation in real-time. This project has demonstrated the effectiveness of this approach through the development and evaluation of a deep learning model trained on a task-specific dataset. The successful implementation of this project offers numerous opportunities for applications in various domains such as computer vision, natural language processing, and content creation. Further research in this area could lead to significant advancements in the field of AI and machine learning, ultimately leading to more accurate and efficient methods of generating images from textual descriptions.



THANK YOU