# AUDIO TO IMAGE GENERATION USING STABLE DIFFUSION

Thesis by

ABINANDA E S

223201

In Partial Fulfillment of the Requirements for the

Degree of

Master of Science in Data Analytics and Computational Science


SCHOOL OF DIGITAL SCIENCES

KERALA UNIVERSITY OF DIGITAL SCIENCE, INNOVATION AND TECHNOLOGY

Trivandrum, Kerala

Supervisors: Dr. ASWIN V.S

2024

ABINANDA E S
223201
ORCID: [Author ORCID]

# ACKNOWLEDGEMENTS

I want to express my deep gratitude to Dr. Aswin V S, Professor, School of Digital Sciences at Digital University Kerala, Trivandrum, for his invaluable guidance and mentorship, which were instrumental in the successful completion of this project. I extend my appreciation to Prof. Saji Gopinath, Vice Chancellor, of Digital University Kerala, for granting access to the university's facilities and resources. Furthermore, I am deeply grateful to my friends and family for their steadfast support, inspiration, and aid throughout the project's journey.

# ABSTRACT

The project explores the new dimensions of interaction and creativity in generative diffusion models. The audio-to-image generation project is built on next-gen Stable Diffusion models and generates images from sound input, unlocking interior design, art, therapy, and education opportunities. The work explored in this project underscores the potential of AI and the transformation it can bring, leading the way for innovations and applications. This system works using a 2 step process: audio-to-text conversion and then using that text, text-to-image generation. By using natural language processing (NLP) and a pre-trained Stable Diffusion model, we set up a smooth pipeline that converts recorded or uploaded audio into high-quality visualized images relevant to the audio.

The system is composed of an interactive user interface to allow the users to upload the pre-recorded audio files, or record the descriptions directly. The interface also lets users type their text prompts if they want to. After taking in the input, the system processes the audio to get the text representations that are then passed into the Stable Diffusion model to generate the final image.

# TABLE OF CONTENTS

# LIST OF ILLUSTRATIONS

# LIST OF TABLES

*Chapter 1*

# INTRODUCTION

## 1.1 Introduction

Generative models are one of the essential components of artificial intelligence (AI). The aim is to understand how data is generated from the distribution. Generative models have numerous applications and they can be analysed in many ways[1]. Diffusion models are generative models used to produce new data that closely approximates the training data, which is used to generate images, videos, and text. The application of generative diffusion models has yielded remarkable outcomes in text-to-image generation. The generation of images is done by using a pre-trained model. The model generates images or data we need, similar to the data they already trained. In the case of image generation, diffusion models work by adding noise to the data and then learning to remove the noise from the data to generate realistic images[2]. The model will apply this denoising algorithm in every seed to generate the desired images[3][4].

Over the past few years, generative models have become advanced, by enabling the creation of human-like natural language, generating high-quality images. This also includes the generation of images from the text prompts. Diffusion models have become the most recent advancement in deep model generation at the highest level. Diffusion models are also known as diffusion probabilistic models. The initial diffusion model was established in 2015 at Stanford by Jascha Sohl-Dickstein, a research scientist in Google's Brain group. Diffusion models have connections with different generative models[3][5].

| Models | Year |
|---|---|
| VAE | 2013 |
| GAN | 2014 |
| Normalising Flow | 2014 |
| Autoregressive model | 2016 |
| Energy-based model | 2000 - 10 |

Table 1.1: Generative models

[3]

Each of these models offers a unique approach to the generation of data. VAE (Variational Autoencoders) This model learns how to encoder and decoder input data to the latent space by mapping it. Both diffusion and VAEs are probabilistic models, which generate data using latent distribution into complex data distribution. Generative Adversarial Networks (GANs)[6] Consist of a two-part Generator and discriminator. Which generator will generate images that are similar to the original one and the discriminator will classify it by whether it is a fake or real image. The main disadvantage of GANs is instability in the training process which is caused by non-overlapping between the input and generated data. One of the solutions is injecting noise into the discriminator with the suitable noise by a diffusion model. In Normalising Flows which generate tractable likelihood estimation. where by using diffusion models it offers a straightforward, scalable approach to high-dimensional data. The autoregressive model works by the product of conditional distributions to improve the data quality the distribution is smoothed by using a smooth distribution, then the distribution is learned by the autoregressive model after that the distribution is denoised by applying the denoising approach. Finally, although EBMs offer a versatile method for modeling energy landscapes, diffusion models utilize a systematic noise process that improves training stability and increases sample diversity [7][1].

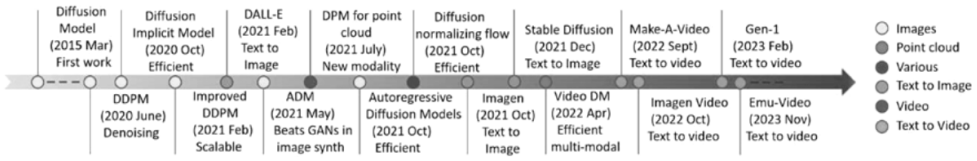


Figure 1.1: History of diffusion models [8]

Here is the chronological summary of important progress in the field of diffusion modeling. In March 2015 the concept of diffusion model was introduced. Then it makes remarkable developments in the generation of technology. The first denoising model Denoising Diffusion Probabilistic Model (DDPM)[9][10] was introduced in 2020 June. It is a Markov chain that is trained by variational inference and it produces high-quality images. In October 2020 the diffusion implicit model was invented, to improve the efficiency of the diffusion process. In February 2021 the first version of DALL-E was developed by OpenAI, which was a text-to-image generation model. After that, the next version of DALL-E developed, DALLE-2. Then in October 2023, DALLE-3 was developed. Adaptive Diffusion Models were introduced in

2021 in May. ADM (Adaptive Diffusion Models) outperformed GANs[11] in image synthesis, demonstrating the superiority of diffusion models in this area. July 2021 the Diffusion models were adapted for point cloud data, expanding their use to new modalities. In October 2021 Multiple advancements, including Improved DDPM, Autoregressive Diffusion Models, and Diffusion Normalising Flow, enhanced scalability and efficiency. Additionally, Imagen was introduced, which is another text-to-image generation model. Stable Diffusion became another prominent text-to-image model in December 2021. Which is one of the best models that recently developed for image generation. Then in April 2022 introduction of Video DM (Diffusion Model), in April 2022, the introduction of Video DM (Diffusion Model) marked a significant milestone in the utilization of diffusion models for video data processing and analysis. In September 2022 Make-A-Video demonstrated text-to-video generation, pushing the boundaries of diffusion models. October 2022 the Imagen Video extended the text-to-video capabilities, emphasizing high-quality video synthesis from textual descriptions. Gen-1 became further advanced in text-to-video applications in 2023 February, continuing the trend of innovation. In November 2023 Emu-Video, another text-to-video model, highlighted the ongoing development and refinement in this area. Research and studies are going on diffusion models for further development in the field of Artificial Intelligence[12][6].

Diffusion models have 3 different categories: Denoising Diffusion Probabilistic Models (DDPMs), Score-based Generative Models (SGMs), and Stochastic Differential Equations (Score SDEs). DDPMs are the main denoising model which uses Markov chains trained by various sample data. During the training, it makes good knowledge of the relation between the clean data and the noisy data with which it converts the noise into the desired image. In SGMs it works by using score-based training, it will estimate the score for all noisy data by training the model. Stochastic processes are used by score SDEs to predict the evolution of data samples and direct the generative process in the direction of producing high-quality data samples. Some of the popular diffusion models are DALL-E2, DALL-E3, Sora, Stable Diffusion, Midjourney, NAI Diffusion, Imagen, etc. In which we[5].

- DALL-E2: Dall-E 2 was developed by OpenAI in April 2022. Which can produce high-quality images according to the given text.

- DALL-E3: It is the most recently developed model by OpenAI OpenAI's most recent image creation model, DALL-E 3, is a significant improvement

above DALL-E 2. The recent version is integrated into GPT.

- Sora: Sora was developed by OpenAI. This is OpenAI's first text-to-video model. Videos created by Sora are remarkably lifelike and may be produced in 1080p at any resolution for up to one minute.

- Stable Diffusion: Stable Diffusion is created by StabilityAI. Which is one of the major AI image-generating models. This model is remarkable for how well it transforms text prompts into realistic pictures. Its ability to produce images of excellent quality has been acknowledged.

- Midjourney: Midjourney is only accessible via a Discord bot on an official Discord server. Which gives more dream-like art visuals. The recent release of Midjourney is Midjourney v6.

- NAI: As a creative tool, the NovelAI Diffusion allows you to visualize your ideas without limitations and produce graphics like never before. Painting the images in your mind is made possible by this.

- Imagen: This model was developed by Google. Which is a text-to-image model. It is popular for its photorealism and deep language understanding. It generates high-fidelity images and encodes text using large transformer language models.

Diffusion models have various applications across different fields because of their capacity to produce high-quality samples and predict complex data distributions. These models are mainly used to generate images or videos from the given text or data. These are also applicable in domains like Education and Personalized Learning, Interior design field, Marketing and Advertising, Entertainment and Media, Graphic design, Fashion and Retail.

### 1.2 New contributions

This report mainly explores the application of diffusion models in the field of interior design, focusing primarily on the use of the pre-trained Stable Diffusion model for audio-to-image generation. We investigate how Stable Diffusion can assist the interior design domain by generating realistic images based on specific

audio and text prompts, providing designers with a powerful tool to visualize and create detailed, accurate interior design concepts. A stable diffusion model has many benefits over a diffusion model, greater sample quality, improved training stability, and improved flexibility in conditioning on new information. Because of these characteristics stable diffusion is a successful tool for image generation, text-to-image generation, and image editing.

The problem addressed in this project is the development of a diffusion model-based system for the generation of images from the audio or textual prompt by the user. The system works by converting the audio input into a textual prompt, and then from the text it will convert into the desired image using the pre-trained stable diffusion model. Users can either upload or record audio and can also give the textual prompt for image generation.[13]

*Chapter 2*

# MATERIALS AND METHODS

This project uses speech recognition, text-to-image generation and a user-friendly interface to convert audio prompts into visual images. The methodology involves several steps, each using specific technologies and frameworks to get efficient and good results.

The first step is to initialize the speech recognizer. The speech-recognition library is used for this, with the Recognizer class instantiated to handle audio input and transcription. This is the part where spoken words are converted into text. Users have two options to provide audio input: upload pre-recorded audio files or record new audio within the application. For uploading, Streamlit provides a file uploader component that supports common audio formats like WAV and MP3. For recording new audio, the system command arecord is used to capture audio input from the system microphone, and save it as an audio file for further processing. This dual option ensures flexibility and convenience for the user.

Once the audio file is uploaded or recorded it gets processed to extract the spoken words. The audio-to-text function does this by loading the audio file and using Google's Speech Recognition API to transcribe the audio into text. This function has robust error handling to handle scenarios where the audio can't be understood or there is an issue with the recognition service. So the app can reliably convert spoken language into text prompts which is needed for the next stage of image generation.

The transcribed text is now used as a prompt to generate an image. The project uses the "stabilityai/sdxl-turbo" model from the diffusers library, a text-to-image model developed by Stability AI. This model is pre-trained on huge datasets so it can generate high-quality images from text. The model uses diffusion techniques to generate images that are both realistic and imaginative. It's loaded and run on a device based on the availability of CUDA for optimal performance. The image generation is guided by the prompt and the output image is generated with a specified guidance scale to improve the output and relevance to the input prompt.

Streamlit is used to present and interact with the user. The web interface allows the user to upload or record audio, see the transcribed text and generate images based on the prompt. Once the image is generated it is shown in the app and the user

can download the image or save it to a specific file path. PIL is used for image file operations so the image can be saved in the desired format and location. This interface is designed to be easy to use and accessible from audio input to image output.

Overall, this project integrates multiple advanced technologies to create a robust and user-friendly application that bridges the gap between audio input and visual output. By leveraging the capabilities of advanced speech recognition and the cutting-edge "stabilityai/sdxl-turbo" text-to-image model, the application provides high-quality image generation based on audio prompts. This innovative approach demonstrates the potential of combining different AI technologies to create solutions that enhance user interaction and experience, offering a novel and engaging way to transform spoken words into visual art.

*Chapter 3*

# EXPERIMENTAL ANALYSIS

## 3.1 Working of Diffusion Models

Diffusion works in a dual-phase mechanism. That is, it mainly works in 2 step process, a Forward diffusion process and a reverse denoising process. The diffusion model generates data using the "reverse diffusion" concept. It is running on the principle of Iteratively improving a noise construct to get quality samples such as images by using this diffusion model. A random noise vector, usually taken from a simple distribution like Gaussian, is used to start this process. This noise vector is what the generation begins with. The model will then apply many transformations to this noise vector by a neural network called the diffusion process[12].

At every stage in the diffusion process, the model lessens the noise in the input and maintains key properties of the target distribution, typically the training data distribution. This is accomplished by varying levels of noise at every step during generation, reducing noise intensity as the generation process continues. The model trains to transmute this noise vector into a useful sample reflecting the data distribution.

The most important thing diffusion models can do is sample well, that is, when we need a bunch of samples from the true data distribution, a diffusion model can generate many high-quality samples by simply introducing more or less noise at each step.

**Step-by-step diffusion process:**

Data preprocessing: Preprocessing the data is the first step in ensuring correct scaling. Standardization is typically used to transform the data into a distribution with a variance of one and a mean of zero. This gets the data ready for changes that will come later in the diffusion process. This process also involves data cleaning, data normalization, and data augmentation to increase dataset diversity, especially in the case of image data[4].

Forward Diffusion process: In this process, the model starts with a sample from a distribution, typically Gaussian distribution. Then "diffuse" the sample by applying a sequence of transformations which is noise. Which is a Markov chain of diffusion

steps in which we randomly apply noise to the original data.

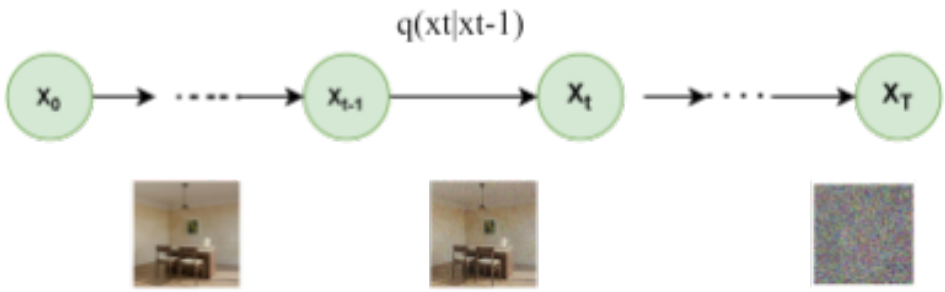


Figure 3.1: Forward Diffusion Process

Reverse diffusion process: In this process, the model learns to remove the noise from the data and generates realistic data. It involves neural network learning to remove noise from an image. The reverse diffusion method involves denoising the data by the different noise patterns that are added at each stage. This is not an easy operation; rather, it requires complex reconstruction. The challenge of turning some random noise into a meaningful image is difficult. The model predicts the noise at each stage using the knowledge it has learned, and then it removes it with care[12].

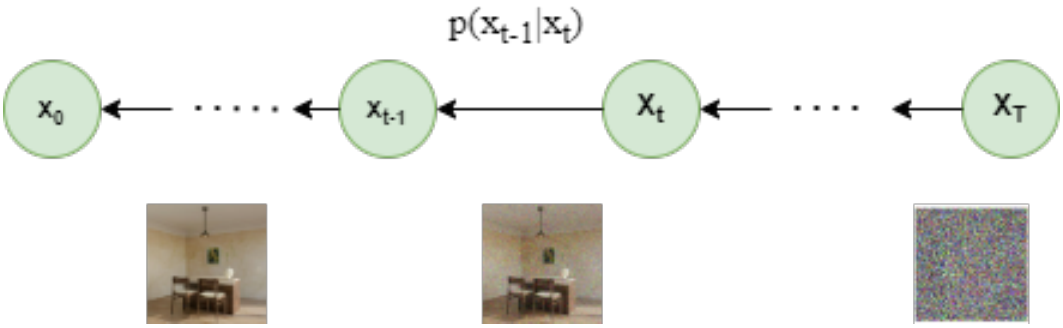


Figure 3.2: Reverse denoising process

Operating on top of diffusion models, Stable Diffusion is a development of these models with the main aim of alleviating issues associated with training stability and sample quality. It extends the tradition of diffusion models but offers several innovations that make it more powerful and robust. Stable Diffusion aims to stabilize the training dynamics and make it easier to converge. This is done by supervising

specific strategies for the diffusion process and using architectural improvements. As an example, Stable Diffusion, a diffusion process that we hand-designed to prevent the vanishing or exploding gradient problems we encounter when training deep neural networks resulted in a more stable gradient[14][15].

### 3.2 Architecture Of Stable Difffusion

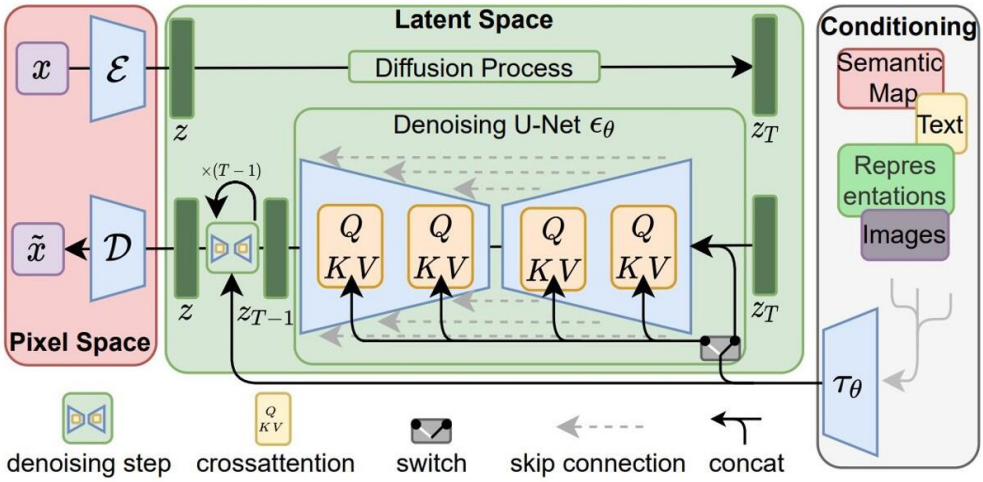


Figure 3.3: Architecture of Stable Diffusion

[16]

Stable diffusion architecture consists of three main components, One for text processing and the other two for lowering the sample to a lower dimensions latent space and then denoising random Gaussian noise. The stable diffusion model is a latent diffusion model. It works on the latent space the other diffusion models work on pixel space. Diffusion in latent space is faster than compared with the pixel space. The other models like Imagen and DALL-E are working on pixel space. The compression of the image in the latent space is done by using the technique Variational Autoencoder[17].

The stable diffusion architecture works on a high-level text prompt that is encoded by a text encoder and fed into a diffusion model along with a noise patch then the generated image representation is decoded into an image. The diffusion process in image or pixel space is computationally slow, even for small images, there are too many pixel dimensions to deal with. Thus in the stable diffusion process, we compress the image and represent it in a lower dimension space called latent space, which would make the diffusion process computationally efficient and this is done

using a separate neural network called variational Autoencoder. Which is made up of two parts: an encoder and a decoder. The encoder compresses an image to a lower dimensional representation in the latent space the decoder restores the image into the pixel space from the latent space. The variational autoencoder is itself trained independently.

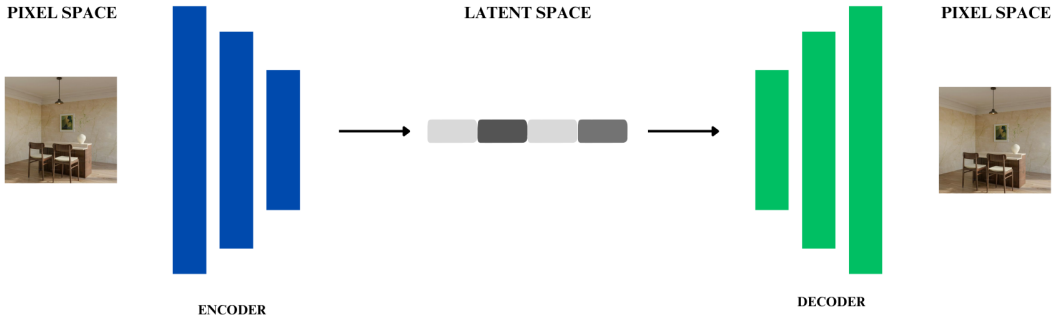


Figure 3.4: Variational Autoencoder(VAE)

Once trained its weights are frozen so the auto encoders ensure that pixel images can be converted to and from the latent space. The diffusion process would take place in the latent space therefore the model is called the latent stable diffusion model.

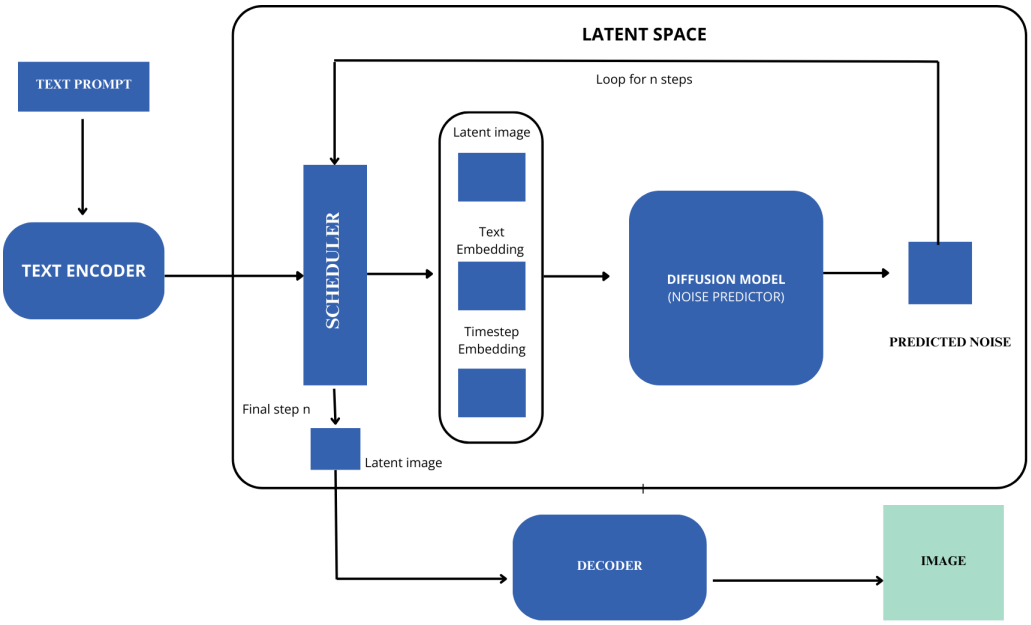


Figure 3.5: Latent Space

The denoising process takes place over a series of time steps. Several time steps are predetermined therefore the diffusion model is invoked during each time step or the

loop. It is the scheduler that orchestrates this flow. The inputs to the noise predictor during each time step are image latent, text embedding, and time step embedding. During inference in the very first step latent image is a randomly generated noise patch. In subsequent steps, the latent image is prepared by denoising the previous image based on predicted noise. The text embedding was provided by the text encoder. The time step embedding represents the noise level in the current step, this helps the diffusion model predict the noise in the input latent image at the final step you get a denoised image which is converted to a pixel image by the decoder.

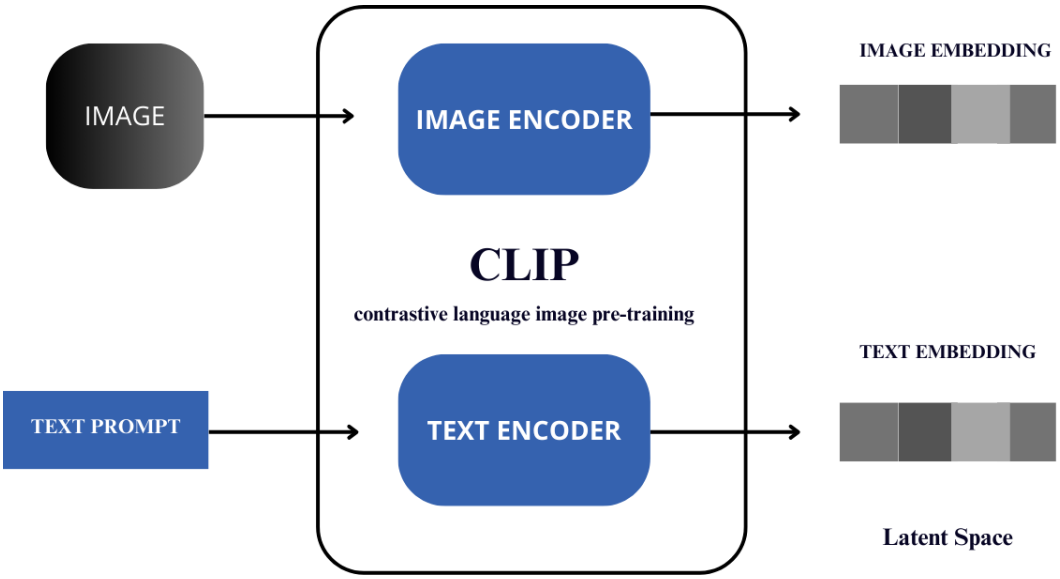


Figure 3.6: CLIP

During image generation starting from a random noise image, how would the model know what image to generate? We want to generate images based on text prompts and not a random image. So we must find a way to incorporate the text prompt in the image generation process. For this, we need to create a latent representation of the text prompt. First, it is done by a text-encode. This text encoding is carried out by a model known as a clip or contrastive language image pre-training which is made up of both text encoder and image encoder the model is pre-trained on a large data set that includes both images and corresponding textual descriptions enabling it to understand the relationships between Visual and textual information. This model is pre-trained independently. Clip is not a generative model and is not used for image generation. This is how the text encoder in the clip generates a text embedding in latent space, which is then passed to the diffusion model and this text embedding is used to condition the image generation process in stable diffusion.

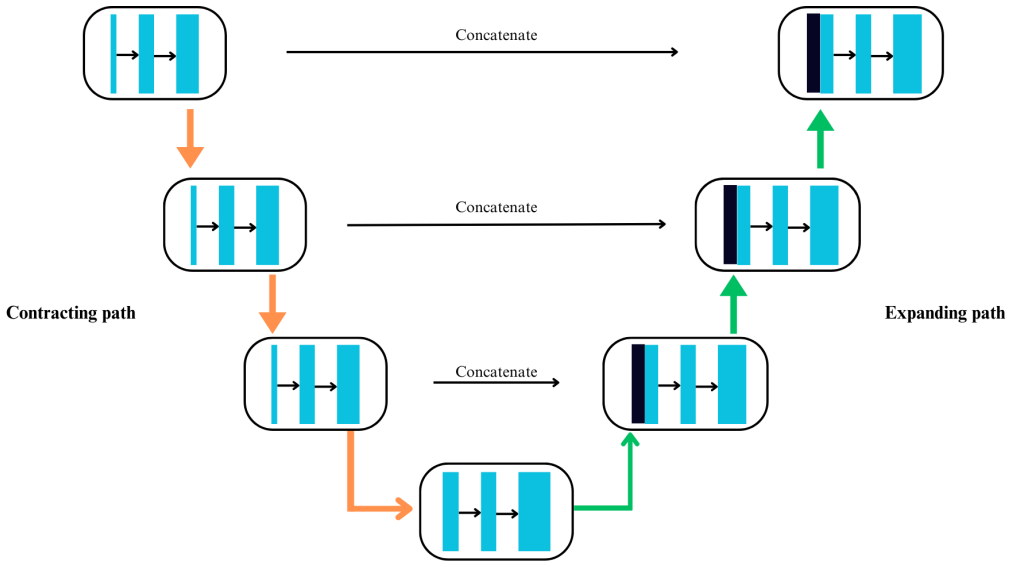


Figure 3.7: U-net

The diffusion model uses U-Net, a neural network architecture as a noise predictor. U-Net is a neural network architecture with a Contracting and Expanding path and it is named after its u-shaped architecture. The Contracting path is made up of many blocks where each block typically consists of convolutional layers followed by a Max pooling layer. The purpose of this path is to gradually reduce the spatial dimensions of the input image while extracting an increasing number of features. The expanding path is made up of blocks that have an upsampling layer followed by convolutional layers. Its role is to reconstruct the image from the features extracted by the Contracting path and restore it to its original Dimensions. The skip connections are designed to connect corresponding blocks between the Contracting and expanding paths. This allows image information from earlier stages of the network to directly flow to later stages thereby preventing information loss due to contraction and expansion[17].

We still need to condition this U-Net Network so that we can generate images based on text. For this attention layers are added to the U-Net architecture. Text embedding and time step embeddings are directly integrated across multiple layers through a cross-attention mechanism. This conditions the denoising process. Convolutional layers in U-Net primarily learn image features. While attention layers help in incorporating textual information into the image-generation process. The combination of these layers allows the model to generate images based on textual descriptions, by leveraging both visual and semantic information and this results in the generation

of images based on textual descriptions and this is how text-to-image generation works.

**Project WorkFlow**

The project focuses on the generation of images from the audio and text prompt using a stable diffusion model. Generation of the image is a two-step process: 1) Audio-to-text conversion and 2) Text Image generation using stable diffusion.

Here is the workflow of the project:

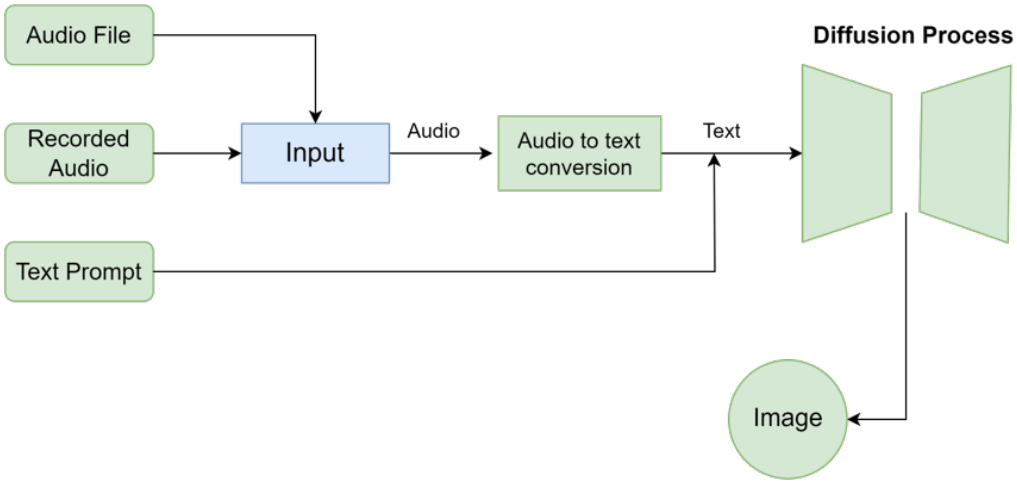


Figure 3.8: Work Flow

The task consists of multiple stages to generate images from the audio input, by using very advanced models both for audio recognition and image generation.

Step 1: Audio Uploading or Recording

Step 2: Audio Handling and Transcription

Step 3: Generating Image from Text

Step 4: User interface setup

Step 5: Display and Manage Generated Image

**Step 1: Audio Uploading or Recording**

This process starts with uploading a pre-recorded audio file or the user can record audio directly on the system's interface. While receiving the audio input, the system uses Python libraries and machine learning models to achieve the desired outcome.

Then the audio file will be speech-recognized and converted into written text. This is done through a more advanced model offered by Google in their Speech Recognition API, specifically trained to convert human speech into text with a very high level of precision.

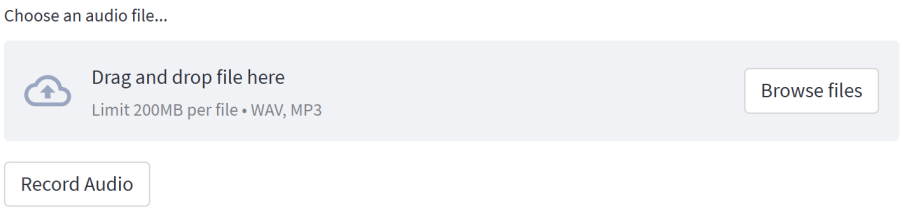


Figure 3.9: Audio input

### Step 2: Audio Handling and Transcription

First, we are using the speech-recognition library which deals with audio transcription to text. It offers an easy and simple-to-use API to perform any speech recognition task. We are using Google's speech recognition API in this project, as it is very accurate and robust in converting audio files to text. We first load the audio using pydub which provides a simple library for manipulating audio files and provides support for many more file formats. This audio is then pre-processed in a format that can be used for further processing and the speech-recognition library takes this audio data and transcribes it into text.

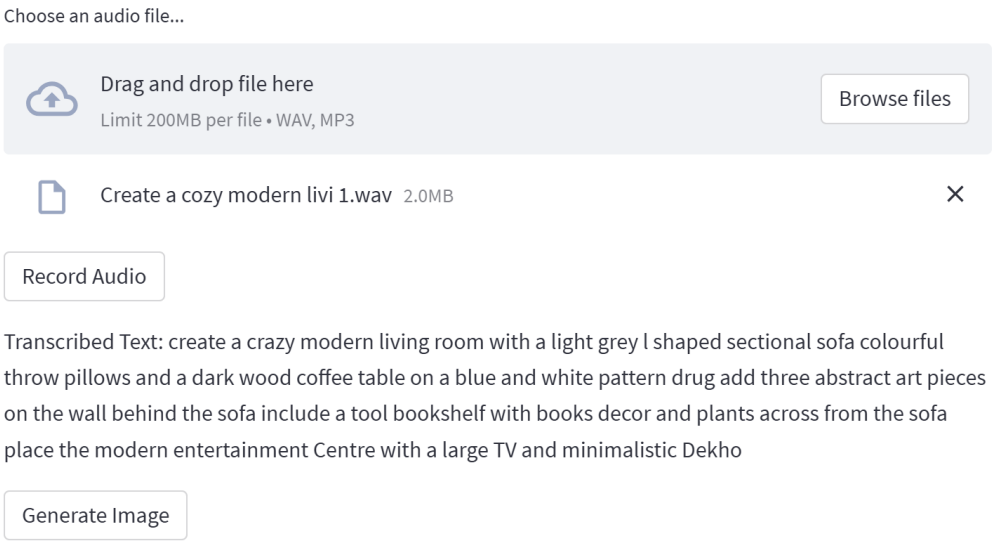


Figure 3.10: Transcribed Text

**Step 3: Generating Image from Text**

In this stage the image generation takes place using the stable diffusion model. At the stage of image generation, the system feeds the textual prompt to the Stable Diffusion model. This model analyses the textual description and produces an image based on what it reads. The generated image is the representation of the transmitted prompt from the audio file. The image generation is using the diffusion library, Stable Diffusion is in the diffusers library by Hugging Face. The stability model used in this project is , stabilityai/sdxl-turbo, this mosaic model is great for generating detailed and photorealistic images from textual descriptions.

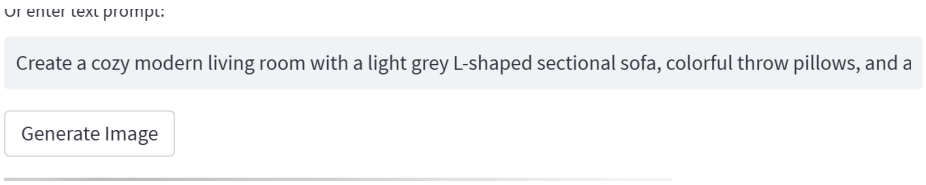


Figure 3.11: Text prompt

**Step 4: User interface setup**

The entire system works on an user-friendly interface Streamlit.A web-based application that enables users to upload audio files, user can record audio instantly and manually input text prompt, and it will display the corresponding image for the user input. The streamlit display contains options for uploading audio, recording audio, entering text prompts, and generating images.

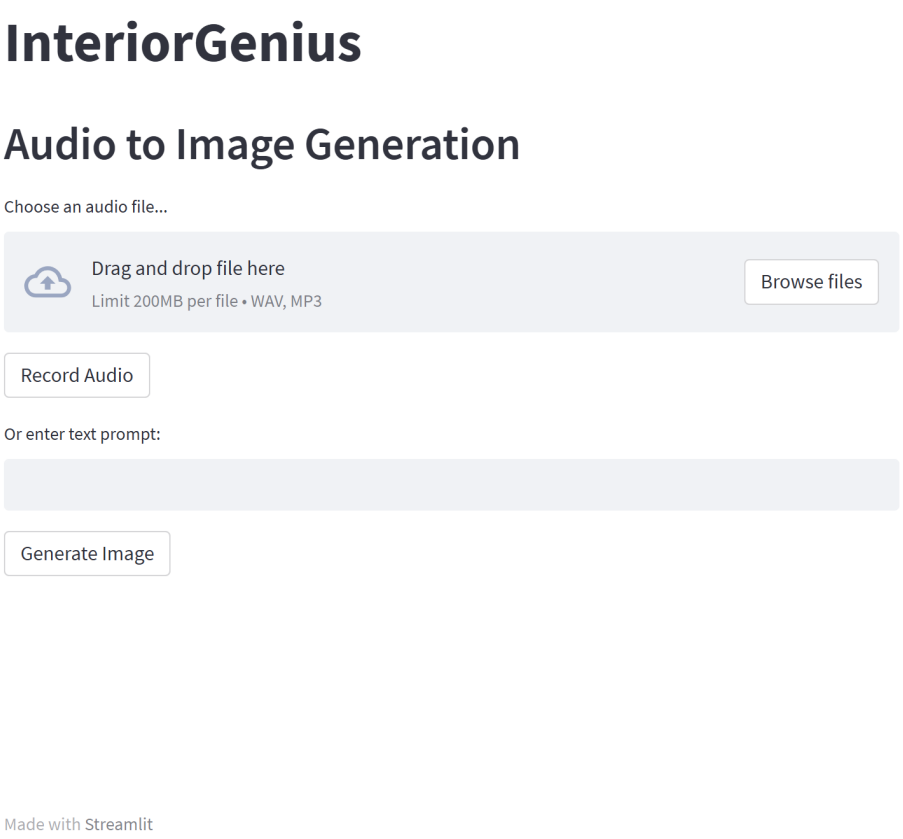


Figure 3.12: Streamlit Interface

**Step 5: Display and Manage Generated Image**

After generating an image it will display on the streamlit interface with a caption that we provided. The application also provide the facility for user to download and save the generated image in the specified file path.

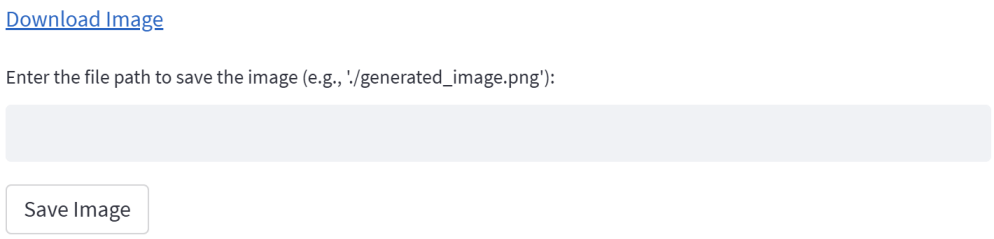


Figure 3.13: Display

*Chapter 4*

# RESULTS AND DISCUSSIONS

The results of our project showcase a significant achievement in the area of the application of stable diffusion in the field of interior design, which is mainly focused on the generation of images from textual description. We evaluated our model by using various text prompts which described the interior model which fed into the stable diffusion model to generate the corresponding images. Also, the quality of the images generated was generally of a high standard, whereby the majority of the images were good and stylistically more attractive. That is why Stable Diffusion is beneficial for an interior designer who needs to make a schematic vision for a client or to experiment with the positions of the objects in the room.

In this text-to-image generation model, images are generated using a pre-trained model SDXL-Turbo developed by Stabilityai. Which is a generative text-to-image model based on the novel training method called Adversarial Diffusion Distillation. It enables high-quality image sampling of huge-scale foundational image diffusion models in 1- 4 stages. To guarantee good picture accuracy, particularly in the low-step domain of one or two sampling steps, our method combines an adversarial loss with score distillation to use large-scale standard image diffusion models as an instructional input. With just one network assessment, the quick generative text-to-image model SDXL-Turbo can create realistic-looking images from a text prompt.

The Streamlit app InteriorGenius consists of a space for uploading the audio file or for recording the audio and for entering the user textual prompt, a run button, and a space that displays the image corresponding to the user input prompt.

# InteriorGenius

## Audio to Image Generation

Choose an audio file...

Drag and drop file here
Limit 200MB per file • WAV, MP3
Browse files

Create a cozy modern livi 1.wav 2.0MB

Record Audio

Transcribed Text: create a crazy modern living room with a light grey l shaped sectional sofa colourful throw pillows and a dark wood coffee table on a blue and white pattern drug add three abstract art pieces on the wall behind the sofa include a tool bookshelf with books decor and plants across from the sofa place the modern entertainment Centre with a large TV and minimalistic Dekho

Generate Image

Generated Image

Download Image

Enter the file path to save the image (e.g., './generated_image.png'):

Save Image

Figure 4.1: Result 1

This is how the Audio image generation using stable diffusion generates images for the given audio file. This is the output image for the uploaded audio which transcribes into “Create a cozy modern living room with a light grey L-shaped sectional sofa, colorful throw pillows, and a dark wood coffee table on a blue and white patterned rug. Add three abstract art pieces on the wall behind the sofa. Include a tall bookshelf with books, decor, and plants. Across from the sofa, place a modern entertainment center with a large TV and minimalistic decor.” Also, there is an option for downloading the generated image for future usage.

Here is another output image when we give any textual prompt as the input for the image generation.

Or enter text prompt:

Create a cozy modern living room with a light grey L-shaped sectional sofa, colorful throw pillows, and a

Generate Image

Generated Image

Download Image

Enter the file path to save the image (e.g., './generated_image.png'):

Save Image

Figure 4.2: Result 2

*Chapter 5*

# FUTURE SCOPE

In general, the Audio to Image generating system is for making auditory content accessible to those with hearing impairments, by converting audio information into visual formats. The use of audio-based image generation in hearing-impaired individuals has the potential to make substantially more things accessible and comfortable in how they live. One such application is developing a system that can translate ambient sounds to be visualized on-screen for real-time sound visualization. The doorbell rings, alarms, and even a baby's crying sound can be retransmitted as visual signals projected on walls or onto smart devices or lighting systems to provide important auditory alerts for people with hearing problems. This system makes their environments more responsive, accommodating, and supportive of their needs.

Also, by using this system one can personalize their design without the help of any designers. Using audio recordings, clients can convey their vision for a room or space, which we enhance into detailed visual mock-ups. Helping designers visualize and experiment with concepts from verbal descriptions, speeding up the concept briefing phase, and improving communications with client requirements greatly.

There are many other fields, including educational and communication enhancements, Marketing and Advertising, Virtual Reality (VR), and augmented Reality (AR) which utilize this technology. Exploring these future applications could significantly expand the scope of audio image generation using the static diffusion model, provide new solutions, and enhance various industries by integrating audio and visual technologies.

*Chapter 6*

# LIMITATIONS

There are also some limitations to using diffusion models. There may be Distortion when the number of subjects that we are given exceeds. Sometimes they are poor at producing text within images. Diffusion models are computationally expensive while it will take more memory to train the data. The existing model is good in many dimensions, but there are still limitations, especially when dealing with prompts that ask for a very detailed, specific answer. Also, the conversion of audio to text is a complex process, there may be misidentification of sounds. The process of integrating the audio-to-image generation system with the already existing smart home devices and infrastructure might have compatibility issues which may be addressed by additional hardware or software changes In the future, more developed versions to include more technical detail and personalization, integration with augmented and virtual reality visualization and ethical considerations of its use will broaden the potential scope and impact of this technology. This allows other AI-based interior design innovations to build off this project, as well as supporting efficient, creative, equitable design processes of the future.

*Chapter 7*

# CONCLUSION AND REMARKS

Thus, by concluding that there is enough potential for Stable Diffusion in the area of interior design image generation from text descriptions. In general, it is best suited for generating natural and visually appealing pictures. Further studies could be directed at developing this approach to make its generative mechanisms even more precise and utilize the model in broader design processes. When it comes to an interface, Streamlit has been effective in allowing users to engage with the technology and it allows users to give their prompt corresponding to their desired ideas about the interior.

Stable Diffusion's application scope and performance for audio-based interior design image generation are promising, which suggests many possible directions for future work to better deploy it and increase its efficacy for practical applications. One of the important features is the model´s ability to manipulate very complex and particular prompts. This might include larger and more diverse datasets and an integration of more advanced natural language processing (NLP) methods that improve the trained models' linguistic competence. Including the user-specific customization and personalization features can improve the model's user experience due to the possibility for the model to improve the outputs based on user feedback and preferences.

# REFERENCES


[1] Stanley Jothiraj and Fiona Victoria. "Phoenix: Federated Learning for Generative Diffusion Model". PhD thesis. 2023.

[2] Shrishti Shah et al. "Generative AI for Text to Image: A Comprehensive Survey". In: *Making Art With Generative AI Tools*. 2024, pp. 17–44.

[3] Ling Yang et al. "Diffusion models: A comprehensive survey of methods and applications". In: *ACM Computing Surveys* 56.4 (2023), pp. 1–39.

[4] Jascha Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning*. 2015, pp. 2256–2265.

[5] Chenshuang Zhang et al. "Text-to-image diffusion model in generative AI: A survey". In: *arXiv preprint arXiv:2303.07909* (2023).

[6] Calvin Luo. "Understanding Diffusion Models: A Unified Perspective". In: *arXiv preprint arXiv:2208.11970* (2022).

[7] Lucas Theis, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models". In: *arXiv preprint arXiv:1511.01844* (2015).

[8] Xiaolong Wang, Zhijian He, and Xiaojiang Peng. "Artificial-Intelligence-Generated Content with Diffusion Models: A Literature Review". In: *Mathematics* 12.7 (2024), p. 977.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502* (2020).

[11] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

[12] Axel Sauer et al. "Adversarial diffusion distillation". In: *arXiv preprint arXiv:2311.17042* (2023).

[13] Hanting Chen et al. "Pre-trained image processing transformer". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

[14] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[15] M Sasirajan et al. "Image generation with stable diffusion AI". In: (2024).

[16] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[17] Yuanling Zhao et al. “Benggang Extraction Based on Improved U-Net Model from Satellite Remote Sensing Images”. In: *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*. IEEE. 2023.