Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.058456

ARTICLE

Integrating Speech-to-Text for Image Generation Using Generative Adversarial Networks

Smita Mahajan[1], Shilpa Gite[1,2], Biswajeet Pradhan[3,*], Abdullah Alamri[4], Shaunak Inamdar[5], Deva Shriyansh[5], Akshat Ashish Shah[5] and Shruti Agarwal[5]

[1]Artificial Intelligence and Machine Learning Department, Symbiosis Institute of Technology, Pune, 412115, India
[2]Symbiosis Centre of Applied AI (SCAAI), Symbiosis Institute of Technology, Pune, 412115, India
[3]Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia
[4]Department of Geology and Geophysics, College of Science, King Saud University, Riyadh, 11451, Saudi Arabia
[5]Department of Computer Science and Engineering, Symbiosis Institute of Technology, Pune, 412115, India

*Corresponding Author: Biswajeet Pradhan. Email: biswajeet.pradhan@uts.edu.au

Received: 12 September 2024; Accepted: 26 January 2025; Published: 30 May 2025

ABSTRACT: The development of generative architectures has resulted in numerous novel deep-learning models that generate images using text inputs. However, humans naturally use speech for visualization prompts. Therefore, this paper proposes an architecture that integrates speech prompts as input to image-generation Generative Adversarial Networks (GANs) model, leveraging Speech-to-Text translation along with the CLIP + VQGAN model. The proposed method involves translating speech prompts into text, which is then used by the Contrastive Language-Image Pretraining (CLIP) + Vector Quantized Generative Adversarial Network (VQGAN) model to generate images. This paper outlines the steps required to implement such a model and describes in detail the methods used for evaluating the model. The GAN model successfully generates artwork from descriptions using speech and text prompts. Experimental outcomes of synthesized images demonstrate that the proposed methodology can produce beautiful abstract visuals containing elements from the input prompts. The model achieved a Fréchet Inception Distance (FID) score of 28.75, showcasing its capability to produce high-quality and diverse images. The proposed model can find numerous applications in educational, artistic, and design spaces due to its ability to generate images using speech and the distinct abstract artistry of the output images. This capability is demonstrated by giving the model out-of-the-box prompts to generate never-before-seen images with plausible realistic qualities.

KEYWORDS: Generative adversarial networks; speech-to-image translation; visualization; transformers; prompt engineering

1 Introduction

Image generation from speech inputs has been a fascinating research topic but comes with challenges of accuracy of the generated images after translation of audio into pictorial output. It is based on the concept of Generative Adversarial Networks (GANs) which have two different neural networks namely generator and discriminator. These two neural networks continuously improve their performances based on each other's feedback. The generator aims to create realistic content so that the discriminator would be fooled by the newly generated synthetic content [1]. GANs are primarily used in image generation applications but can be expanded in other domains like audio, video, code, and even creative content as well.

Image generation using speech as input is essential and interesting because it more natural interaction between the system and the human [2]. In traditional approaches, text is normally used as an input, but it is not easier, natural interaction when compared with speech. One more advantage of using speech over text is identifying emotions and nuances with speech which gives added expressions and a comprehensive image-generation experience [3].

Natural Language Processing (NLP) has been considered one of the ground-breaking topics in the last few years. With the advancements in transformer architectures, there are magical milestones like GPT3 with billions of parameters and considered as one of the most popular transformers [4]. But translating detailed narratives into realistic pictures is a difficult task. While designs like Stacked Generative Adversarial Networks (StackGAN++) [5] have shown promising results in recent years [6], they have been limited by the visual aspects of their training datasets. More recently, OpenAI introduced a groundbreaking deep network named Contrastive Language-Image Pretraining (CLIP) [7]. The CLIP system comprises two encoders [8], one for text and one for images. Through pretraining on 400,000,000 image-text pairs (consisting of images and their corresponding captions), CLIP learns to generate comparable embeddings for words and pictures that convey related concepts [9]. The versatility of CLIP extends beyond training, making it applicable to visual categorization tasks without the need for further training. CLIP exhibits impressive "zero-shot" capabilities, enabling it to accurately predict entire classes it has never encountered before. When a model attempts to predict a class, it has only been seen once in the training data; this is referred to as "zero-shot learning". Models like CLIP typically excel at zero-shot learning due to their utilization of text information in (image, text) pairs. Even when presented with significantly different images from those in the training set, the CLIP model can often provide a reliable caption for the image [10].

In this research, the authors propose a CLIP-based model designed to generate the best matching image for a text embedding obtained from a speech input as shown in Fig. 1. This input can be used for various purposes, including producing samples for image datasets, snaps of human faces, representative photographs, cartoon or animated characters, image-to-image translation, text-to-image translation, semantic-image-to-photo translation [10], generating photo-realistic paintings using ExGANs [11], face frontal view generation [11], generating new human poses [12], photos to emojis, photograph editing [13], face aging, photo blending [14], super-resolution, photo inpainting, clothing translation, video prediction, and 3D object generation [15]. While current models have embraced a multi-modal approach by incorporating various inputs into their algorithms, the use of speech remains unknown in image generation. A speech-driven generative engine holds numerous promising applications in the realm of virtual reality. However, there has been limited progress in developing a speech-based GAN encoder. The most recent and significant advancement in this direction has been pioneered by Goodfe et al. [16].

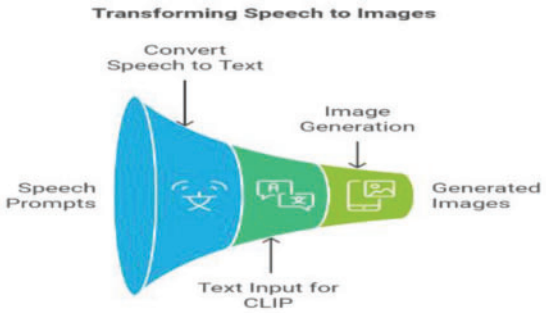

Figure 1: Transforming speech to image

In this paper, a speech-to-image generation model is presented using Vector-Quantized GANs and the CLIP Transformer [17]. The novelty of this approach is to integrate speech as a prompt for generating pictures. These speech descriptions are translated into text embeddings and inputted into the generative model to produce plausible imagery as outputs. The proposed model uses GAN along with the CLIP transformer and Vector Quantized Generative Adversarial Network (VQ-GAN) model, using speech-to-text conversion [18]. The authors aim to develop technology that allows us to visualize human speech, which could have numerous applications in education and artwork [19]. For example, GANs can be used to create various environments and levels in games as well as to add new levels to existing games. The Non-fungible Tokens' (NFT) space, where abstract AI-generated art has seen significant popularity in recent years and gaining attention [20].

CLIP has the potential of cross-modal assimilation useful for text-to-image conversion [21]. It is easier to capture the corresponding descriptions while CLIP models are implemented. On the other hand, VQGAN [Vector Quantized Generative Adversarial Network] enhances resolution and control over detail, creating high-quality and systematic image synthesis by building images as discrete representations [21]. VQGAN allows for high-quality and structured image synthesis by rendering images as discrete representations, improving resolution and detail control. CLIP + VQGAN can effectively align text with image features due to CLIP's pre-trained language-vision model, transcending conventional GANs (like StackGAN and AttnGAN) floundering with subtle language cues. Earlier models sometimes miss minor textual features, particularly for conceptual descriptions or complicated settings, as they depend on the latent space method of CLIP. With excellent control over subtle details, VQGAN's quantization technique produces high-quality, high-resolution images comparable to or better than StackGAN++. The resolution capabilities of prior models, such as AttnGAN and StackGAN++, are restricted. VQGAN is suitable for high-resolution necessities since it creates images with exceptional scaling and structure. CLIP + VQGAN can produce and comprehend visuals more effectively, even with unclear text, due to CLIP's natural strong language understanding. Earlier methods used by DM-GAN did not achieve CLIP's cross-modal understanding depth but added dynamic memory to handle weak text definitions. The pre-trained language processing of CLIP suggests a reliable basis for flexible and coherent outputs. In contrast, previous models need extra modules to adapt to ambiguous or multi-meaning text inputs. CLIP's inherent text-image embedding alignment can reduce computing requirements, making it more effective and versatile than AttnGAN, which extensively uses attention processes. Prior models such as AttnGAN that heavily rely on attention may present computational difficulties, mainly when negotiating with intricate scenes. This prerequisite is reduced, and efficiency is increased by using CLIP. The CLIP + VQGAN model significantly gains flexibility, text-image alignment, and detail quality [21]. Compared to earlier GAN-based models, this model is especially effective for applications that need computing efficiency, high-resolution imagery, and subtle language interpretation.

GANs and NFTs converge on the concept of creativity, with GANs demonstrating creativity through complex algorithms trained on datasets, like how humans draw inspiration from existing artworks [22]. In recent years, recurrent neural networks and GANs have made significant progress in zero-shot recognition and image translation research [23]. The BigGANs are used to generate synthetic photographs that are nearly indistinguishable from the originals. Additionally, StackGANs can be used to translate realistic images based on text-based descriptions.

The novelty of our approach lies in the integration of Speech-to-Text translation with CLIP + VQGAN for image generation. Unlike existing models that primarily use text or image inputs, our model leverages the auditory modality, providing a unique and innovative method for generating images. This technique of generative models is impactful and also enhances their applicability in various fields. While previous works have explored text-to-image generation using GANs and transformer models, the incorporation of speech input as a direct prompt for image synthesis is a novel and unexplored approach. This speech-driven

image generation approach has potential applications in various fields, including education, artwork creation, gaming, and virtual reality environments. The paper showcases significant results and low FID scores of 31 to 65 based on multiple examples of generated images across various topics and domains.

Overall, the main contribution of the paper can be summarized as:

- Integration of CLIP for Text-Image Conversion: The architecture incorporates CLIP to strengthen the alignment between text and image features.
- Novelty in Text-to-Image Generation: This approach introduces a new modality by using speech prompts as inputs for text-to-image generation.
- Combination of Speech Input with CLIP, GAN, and Iterative Latent Space: The model integrates speech inputs with CLIP, GAN, and an iterative latent space process, enhancing the generation process.

This paper delves into an in-depth literature review presenting the history of popular GAN models, transformer-based models, and text-to-image translation models implemented using generative architectures in Sections 2–4. Subsequently, Section 5 highlights the proposed architectural model of this paper, which utilizes speech input prompts for generating images, and the datasets used are discussed. Section 6 is dedicated to discussing the results, and Section 7 states the conclusion and addresses the findings of this concept.

2 Literature Review

GANs involve two deep learning networks locked in a competitive training process. The first network, the generator, acts like a creative artist, trying to produce realistic images. The second, the discriminator, plays the role of a critical art reviewer, aiming to distinguish real images from the generator's creations.

The focused literature review presented here provides context specific to the GAN variants or modifications used, particularly when the work builds upon traditional GAN concepts in innovative ways. The authors have included this review to focus on context, model evolution, and relevance to the study, even though foundational GAN information is widely accessible. This review enables the authors to highlight how these adaptations address specific challenges or limitations in existing GAN models, reinforcing the importance of the chosen methodology. It offers a concise background and prepares readers for the technical details of the methodology and experiments that follow.

2.1 Deep Convolutional Generative Adversarial Network (DCGAN)

Applying the Transposed Convolution Operation on GANs acts as an upscaling operation that helps translate lower-resolution images to higher-resolution images [24]. Conditional Generative Adversarial Network (CGAN) CGAN solves the problem of data generated from noise found in GANs (especially in images with multiple classes) by making sure the generator is creating images only of one particular class [25]. CycleGAN: CycleGAN aims to solve the problem of image-to-image translation. CycleGAN performs the unpaired image-to-image translation. That means the images which are used for training, do not have to represent the same thing [26]. Coupled Generative Adversarial Networks (CoGAN) To confuse the discriminative models, a group of generative models collaborates to synthesize a pair of pictures from two distinct domains. The discriminative models seek to distinguish between pictures derived from the distribution of training data in the relevant domains and those produced from the corresponding generative models [27].

ProGAN Progressive growing of Generative Adversarial Networks One of the issues with GANs can be traced back to instability in Training. The loss of the GAN can occasionally oscillate as the learning of one by the generator and the other by the discriminator is undone. In other cases, the loss may

materialize immediately after the networks converge and the photos begin to appear appalling. ProGAN, or the progressive expansion of generative adversarial networks, is a method for stabilizing GAN training by gradually boosting the produced image's resolution [28]. WGAN (Wasserstein Generative Adversarial Networks): This minimizes an approximation of the Earth-Mover's distance [EM] rather than the Jensen-Shannon divergence as in the original GAN formulation [29]. SAGAN Self-Attention Generative Adversarial Networks: For long-range dependency modeling for image generation tasks, the SelfAttention Generative Adversarial Network, or SAGAN, is used. Only spatially local points in lower-resolution feature maps are used by conventional convolutional GANs to produce high-resolution details. Details in SAGAN may be produced utilizing hints from all feature locations [30]. Additionally, the discriminator may verify the consistency of extremely detailed features in distant areas of the picture [31]. Big Generative Adversarial Networks (BigGAN) The BigGAN method combines a variety of current best practices for training class-conditional pictures while scaling up batch size and model parameter numbers. As a result, photographs with high resolution (big size) and excellent quality (high fidelity) are often produced [32]. Style-based Generative Adversarial Networks (StyleGAN), the Style Generative Adversarial Network, also known as StyleGAN, is an addition to the GAN architecture that suggests significant changes to the generator model. These changes include the use of a mapping network to map points in latent space to an intermediate latent space, the use of the intermediate latent space to control style at each point in the generator model, and the addition of noise as a source of variation at each point in the generator model [33].

2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are used in generating image, video, and voice data. They are trained to utilize two neural network models: a generator that learns to produce new data and a discriminator which distinguishes between real and fake data generated by the generator. The generator aims to deceive the discriminator, while the discriminator possesses all pertinent information. Following training, generative models can be utilized to generate large volumes of data as required [10]. Creating visuals from detailed captions has been a challenge that researchers and developers have previously addressed. However, the introduction of GANs has significantly raised the performance of this task. Designs like StackGAN++ [34] and align (Deep Recurrent Attentive Writer) DRAW [35] have demonstrated promising results in recent years despite being constrained by the visual and textual domains of the training dataset. Additionally, other applications, such as image synthesis and style transfer, have evolved from the two-player methodology introduced by Ian Goodfellow in 2014. More recent Text-to-Image translation applications are implemented using various GAN-based generative models, including Variational Auto Encoder and transformers [36,37]. Latest research papers also discussed stable diffusion models for text-to-image generation with transformer-based architectures as shown in table [38,39]. GAN variants (e.g., ProGAN, StyleGAN) are primarily developed for high-quality image synthesis (e.g., realistic faces, objects), their architectural innovations and techniques have significantly influenced text-to-image generation models. Text-to-image models like DALL·E and Stable Diffusion also require generating high-resolution images. The progressive growing strategy inspired techniques to scale resolution in text-to-image pipelines while maintaining detail. Style GAN's ability to separate content (structure) and style (appearance) has direct implications for text-to-image generation, where the model must interpret a text description and translate it into distinct but coherent visual features [40,41].

3 Working of Stack GAN

Stacked GANS is widely used for producing high-resolution photos with numerous photorealistic features [42]. The text-to-image generation process is mainly divided into two parts [43] as shown in Fig. 2.

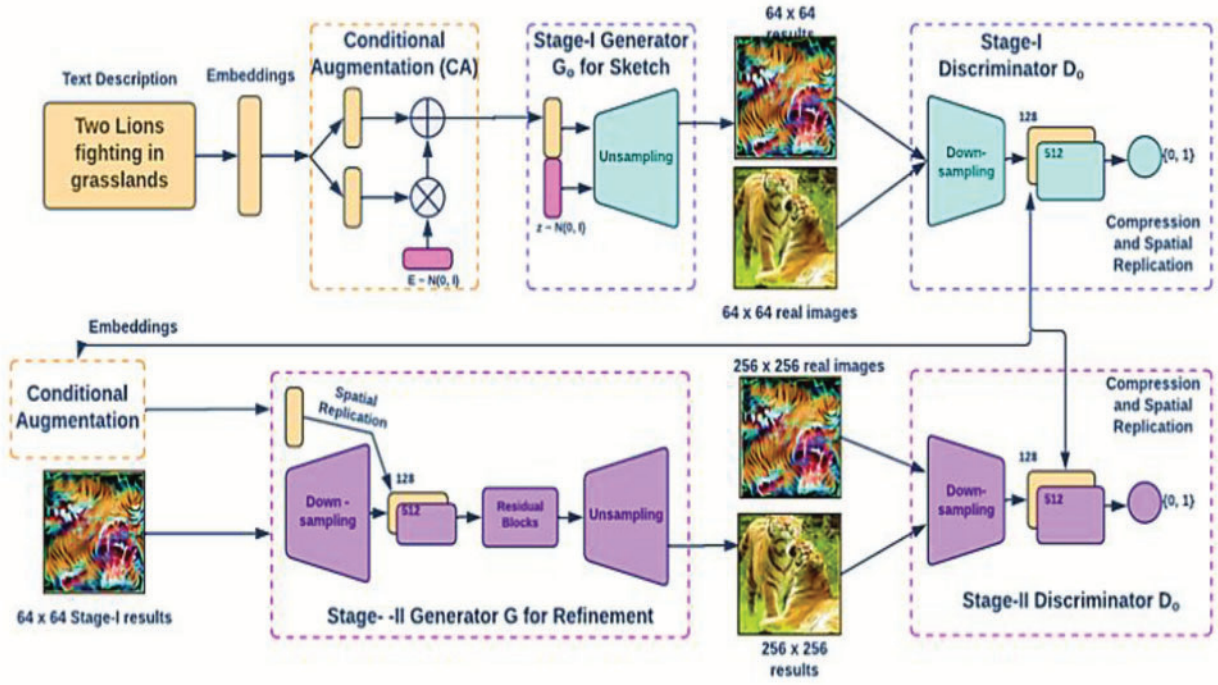

Figure 2: Functioning of stack-generative adversarial networks

3.1 Stage 1 GAN

A basic low-resolution image is created by outlining the basic colors and shape of the object according to the provided text description. Subsequently, the model constructs the background layout using a random noise vector. The GAN then sketches the primitive shape and colors of a scene based on the provided text description, resulting in low-resolution images.

3.2 Stage 2 GAN

It completes the object's details by reevaluating the text description. It rectifies any errors present in the low-resolution image from Stage 1, thereby creating a high-resolution photo-realistic image [44]. This stage utilizes the results from Stage 1 along with the text description as inputs to generate high-resolution images with photo-realistic details. These stages are explored in depth by Li et al. [28].

Functioning of Stack Generative Adversarial Networks StackGAN: This approach utilizes a two-stage GAN system called StackGAN to generate images based on text descriptions [45].

Stage 1: Capturing the Essence

- **Text Encoding**

Translating the textual description (e.g., a sentence describing an object) into a concise representation using techniques like word embeddings [5] or recurrent neural networks (RNNs) [46]. Think of this as capturing the core meaning of the words.

- **Injecting Creativity with Noise**

This encoded text is then combined with random noise, which acts as a spark of creativity for the generator network. Imagine adding some artistic flair to the textual blueprint.

- **Low-Resolution Sketch**

The Stage 1 generator network uses this combined input to create a low-resolution image (e.g., 64 × 64 pixels) that reflects the basic elements from the text description. This is like a rough sketch capturing the initial form.

- **Training the System**

During this stage, the generator learns to translate basic textual information into simple shapes and structures within the image [47]. Simultaneously, a separate network, the discriminator, trains itself to distinguish real images from these initial attempts.

Stage 2: Refining the Details

- **Building on the Foundation**

The next stage combines the low-resolution image from Stage 1 with the original text encoding. This provides the Stage 2 network with both the initial sketch and the textual details.

- **High-Resolution Masterpiece**

The Stage 2 generator network utilizes this information to create a high-resolution image (e.g., 256 × 256 pixels or higher) that closely resembles the described scene.

- **Continuous Improvement**

Similar to Stage 1, both the generator and discriminator in Stage 2 undergo adversarial training [48]. The generator strives to produce increasingly realistic images, while the discriminator refines its ability to discern reality from generated images. In this way, the StackGAN culminates in a high-resolution image that visually aligns with the textual description. It effectively translates the words into a corresponding visual representation.

3.3 Transformer Models

Transformer models were first introduced by Vaswani et al. [49] in the year 2017. The model is mostly utilized for sophisticated natural language processing applications. OpenAI employed transformers to develop its well-known GPT-2 and GPT-3 [50] models. The transformer architecture has developed and branched out into several forms since its introduction in 2017, going beyond language problems to other domains. They have been applied to forecast time series. They are the main technological advancement underpinning DeepMind's protein structure prediction model, Alpha Fold [51]. Transformers serve as the foundation for Codex, an OpenAI source code creation paradigm. Transformers have more recently made their way into the field of computer vision, where they are gradually taking the place of convolutional neural networks (CNN) in a variety of challenging applications; this is attributed to the attention mechanism in transformers capable of memorizing relative positions of features in images [52]. Transformers can still be improved, and researchers are continuously investigating new uses for them.

OpenAI recently unveiled a unique deep neural network that acquires visual concepts through natural language guidance in January 2021. The Contrastive Language Image Pretraining (CLIP) [53] consists of two encoders, one for images and one for text. CLIP's encoders produce comparable embeddings for images and words, capturing similar concepts. When applied to visual classification tasks without training, CLIP can differentiate between objects X and Y in an image dataset by assessing whether the text description "a photo of X" or "a photo of Y" is more likely to be associated with each image.

This transformer has been extensively used in image captioning applications [54,55] due to its ability to create shared representations for image and text prompts. The strong correlation between visual and textual

features makes CLIP a highly practical approach. The model proposed by the authors generates an image whose CLIP embedding is most similar to the given text embedding. Through exploration by a genetic algorithm, the generative network produces an optical image.

4 Generating Images and Videos Using VQGANs

The VQGAN was previously utilized by Ge et al. [15], who generated long-form video using a time-sensitive transformer architecture built on top of 3DVQGAN. This engine was trained on 16-frame clips from benchmark video datasets and effectively produced high-quality videos. The authors used a time-agnostic GAN and encoded temporal information using padding methods between video frames.

Video generation using GANs is not a new concept. Originally designed for image synthesis [37], adapting these methods to video synthesis necessitated encoding the temporal dimension. Some papers have attempted this using RNNs; for instance, MoCoGAN [6] sampled motion vectors to create multiple video frames. More recent methods like MoCoGAN-HD [6] utilize Long Short-Term Memory (LSTM) to assign probabilities to a curve in the abstract multi-dimensional space of a trained picture generator. Tables 1 and 2 highlight the summary of various text-to-image synthesis methods.

Table 1: Summary of text-to-image synthesis methods

Name	Author	Topic	Findings
StackGAN	Zhang et al. [6]	StackGAN: Text to Photorealistic Image Synthesis with Stacked Generative Adversarial Networks	The suggested approach breaks down text-to-image synthesis into a novel process of sketch refinement. Compared to current generative models, this technique produces images with higher quality, more photorealistic features, and greater variety
AttnGAN	Xu et al. [51]	AttnGAN: FineGrained Text to Image Generation with Attentional Generative Adversarial Networks	This study suggests a new multistage attentional generating network for the AttnGAN to produce high-quality images. Additionally, the authors suggest a deep attentional multimodal similarity model to compute the fine-grained image-text matching loss for training the AttnGAN generator

(Continued)

Table 1 (continued)

Name	Author	Topic	Findings
StackGAN++	Zhang et al. [6]	StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks	This paper talks about two versions of Stack GANs, StackGAN-v1 and StackGAN v2. While StackGANv1 successfully generates 256 × 256 resolution images from text, StackGAN-v2 approximates multiscale image distributions and combines conditional and unconditional image distributions jointly.
DM-GAN	Zhu et al. [52]	DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis	The proposed methodology introduces the dynamic memory module for handling ill-defined images. DM-GAN outperforms the state-of-the-art models in terms of both qualitative and quantitative metrics.

Table 2: Summary of text-to-image synthesis methods

Model Type	Strengths	Challenges
Diffusion models	High-quality, realistic outputs	Computationally intensive
GANs	Creative and artistic outputs	Instability during training
Auto-regressive	Pixel-level accuracy	Slow generation process
Retrieval-augmented	Enhanced realism with references	Requires robust retrieval mechanisms
Scene graph-Based	Structured and relationship-aware	Complex preprocessing
Multimodal transformers	Unified learning of text and image	Requires large-scale training data

It's worth noting that the CLIP transformer has historically served as a guidance tool for manipulating images [55]. Recently, Vector Quantized Variational Autoencoder (VQ-VAE) based visual models for video synthesis tasks have been introduced, involving transforming pictures into discrete tokens or frames. This technology has the potential to provide extensive pretraining for text-to image creation. However, in the problem statement addressed in this paper, the authors have implemented a speech-to-image generator. Speech-to-image translation is an extremely complex task involving the creation of a speech embedding network and generating paired probabilities of parts-of-spoken-word-to-image pairs [40].

Generating images from speech input introduces unique challenges due to the fundamental differences between auditory and visual modalities. Unlike text, which has clear symbolic representations, speech involves nuances like pitch, tone, rhythm, and phonetic variability that must be mapped accurately to visual information. Using speech input for image generation introduces unique challenges that span speech recognition, natural language understanding, and image synthesis. Speech-to-Text Conversion Accuracy-Errors in automatic speech recognition (ASR) can increase incorrect textual inputs, that leads to incorrect

image generation. Its solution can be training on diverse and domain-specific datasets to improve recognition accuracy. Spoken descriptions are often less structured and more ambiguous than written text. Also, variations in tone, accent, pacing, and phrasing can affect the system's interpretation. A possible solution would be to leverage large-scale language models (e.g., GPT) to parse and refine ambiguous speech into structured textual descriptions. Generating images from live speech inputs requires low latency for applications like interactive tools as it should be generated in real-time. Using lightweight, optimized models for ASR and image synthesis (e.g., efficient diffusion models) can make real-time processing possible. Speech systems often struggle with accents, dialects, and noisy environments, affecting downstream tasks like image generation. Training ASR models with diverse datasets that include different accents and dialects will make the model robust. This paper has devised a method to simplify the process of creating complex embeddings. It implements a clever sequence of operations to translate speech into tokenized words and utilizes pre-trained models in conjunction to generate beautiful imagery from speech expressions.

5 Proposed Approach

This study investigates a new modality in text-to-image generation by using speech prompts as inputs, which is novel compared to traditional text prompts. The use of speech allows for more adaptable and natural interactions, which helps applications in accessibility, virtual assistants, and interactive media. The approach is also useful for the users with visual impairments or those in hands-free environments. The model uses CLIP's pre-trained language-image embeddings with a cosine similarity loss function to calculate and optimize the resemblance between the encoded text and image representations. This approach guarantees that generated images align closely with the semantic content of the speech prompt. Using cosine similarity for alignment enhances the interpretability and accuracy of generated images, addressing common issues in generative models where images may only loosely match prompts. Optimizing the latent space, which is critical, allows for finer control over the generated content, enhancing the clarity and fidelity of the final images. The architecture merges CLIP for text-image alignment, a GAN for high-quality image synthesis, and a speech-to-text component to translate spoken prompts into text. This tri-component strategy represents a novel pipeline for multimodal image generation, leveraging the strengths of each model. This approach enables real-time, multimodal interaction where spoken words can directly drive the creation of images. The combination of CLIP and VQGAN facilitates natural language understanding and high-quality image synthesis abilities. The system begins with a random noise or an initial latent vector (image representation), where VQGAN generates an image iteratively. CLIP evaluates the alignment of the generated image with the textual description provided. The system optimizes the generated image by adjusting the latent space in VQGAN, which is driven by CLIP's feedback loop. The study explains using CLIP, highlighting its capacity to generalize across visual domains with minimal training, similar to GPT models. The deployment of CLIP's pre-trained strengths decreases the need for comprehensive dataset-specific training, making the model more efficient and versatile for new applications. The speech input with CLIP, GAN, and iterative latent space optimization, generates better quality and highly realistic images.

CLIP is designed to combine visual and textual understanding in a single model as shown in Fig. 3. Its primary function is understanding and retrieving images based on natural language depictions without requiring specialized training datasets for each task. It utilizes a contrastive learning approach, learning from an extensive dataset of images and their captions. CLIP learns to map images and text into a shared embedding space where similar images and captions are close. CLIP's capability to correspond text to images has made it widely applicable for various AI tasks, including image search, text-based image manipulation, and guiding other generative models (like DALL-E) to produce images aligned with textual descriptions. StackGAN++ is useful for text-to-image generation tasks, mainly where the model generates

images that closely match the input textual description, with more definitive colors and sensible textures than previous GAN-based models. Combining CLIP and StackGAN++ entitles better control over text-to-image generation. CLIP can act as a "guiding model" for StackGAN++, estimating and leading the image generation toward nearer alignment with the input text description. By employing CLIP's capability to diagnose text instructions, it is possible to use a StackGAN++-like model to refine images iteratively, considering the user feedback. While CLIP is good at aligning images and text, StackGAN++ concentrates on forming intricate images from descriptions; combined, they create a robust technique for high-quality, interactive image generation. CLIP's semantic understanding to maintain relevance to the text is its highlight. CLIP's robust semantic understanding can mitigate minor ASR errors in speech-to-text conversion, ensuring the generated images remain relevant. VQGAN + CLIP can handle abstract or vague descriptions often found in spoken language. It also supports real-time applications like interactive art tools. Traditional GANs have limited semantic text-image matching but CLIP + VQGAN can easily generalize the match and produce high quality images. The highlight of our model is it works with unpaired data in the unsupervised learning environment unlike other traditional models. The working mechanism of CLIP + VQGAN begins with the user providing a text prompt, such as "A dragon." VQGAN initializes with a random latent vector z, representing an initial image in its latent space. The text prompt is encoded into a text embedding using CLIP's text encoder, while the image generated from z is encoded by CLIP's image encoder to obtain an image embedding. The cosine similarity between these embeddings serves as a score for how well the image matches the prompt. Through an optimization loop, z is updated via gradient descent to maximize this similarity, with regularization terms ensuring image coherence and quality. The process iterates until the image aligns with the text prompt, and the optimized z is decoded by VQGAN to produce the final image. Thus, combining CLIP for the multimodal understanding with the generative capabilities of VQGAN works very well for textual concepts and visual imagination. Some of the challenges for speech recognition systems include pronunciations, lingoes, background noise, or vague pronunciation, leading to inaccuracies in the transcribed text. For example, Homophones (e.g., "flower" vs. "flour") can be misinterpreted, especially if the context is ambiguous. Another challenge is that the proper nouns, technical jargon, or cultural idioms may need to be recognized correctly during this conversion. Also, spoken language requires more structure than written language (e.g., fillers, pauses, or fragmented sentences). Sometimes, even accurate transcription can deliver incomplete or unclear descriptions (e.g., "a large animal" could mean any large animal, like an elephant, a whale, or a horse).

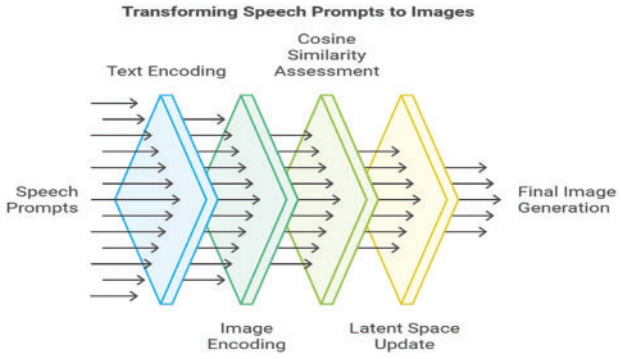

Figure 3: Functioning of stack-generative adversarial networks

Speech usually requires the inclusion of details or prompts necessary for explicit image generation. Speech-based contextual cues like tone or intonation may be misinterpreted during conversion, potentially

losing nuances like emphasis or implied meanings. The Speech may rely on situational or cultural context to convey meaning (e.g., "a ceremonial dress" or "traditional dress" can convey vastly different things based on the territory). Image generation models may need clear, detailed information. The Contrastive Language-Image Pretraining (CLIP) model is a fundamental component in the proposed architecture.

The (CLIP) model aligns image embeddings and text embeddings in a shared latent space. This is integration achieved through a contrastive loss function. The contrastive loss function guarantees that matching text-image pairs are more similar in the latent space than mismatched pairs. The CLIP loss is computed as the negative logarithm of a fraction. The numerator of this fraction is the exponentiated similarity score between the text embedding $\tau\tau$ and the image embedding II, divided by the temperature parameter $\tau\tau$. Compared to non-matching pairs, the loss shows the high similarity between matching text-image pairs, optimizing the model to align text and image embeddings in the shared space.

CLIP model aligns image embeddings and text embeddings in a shared latent space. The contrastive loss function ensures that matching text-image pairs are more similar in the latent space than mismatched pairs. The CLIP loss is calculated as the negative logarithm of a fraction. The numerator of this fraction is the exponentiated similarity score between the text embedding TT and the image embedding II, divided by the temperature parameter $\tau\tau$. Compared to non-matching pairs, the loss indicates the high similarity between matching text-image pairs, optimizing the model to align text and image embeddings in the shared space. The contrastive loss function is mathematically stated as shown in Eq. (1).

$$\mathrm{L\_CLIP} = -\log \sum \mathrm{I}' \exp(\mathrm{sim}(\mathrm{T}, \mathrm{I}')/\tau) \exp(\mathrm{sim}(\mathrm{T}, \mathrm{I})/\tau) \tag{1}$$

where:

Similarity Function (sim(T, I) sim(T, I)): Measures the similarity between the text embedding TT and the image embedding II. Similarity Function is typically implemented as the dot product of normalized embeddings.

Temperature Parameter ($\tau\tau$): The Temperature Parameter is a learnable or fixed scalar value that controls the sharpness of the probability distribution.

The numerator represents the similarity between the text TT and its corresponding image II, scaled by the temperature parameter. The denominator states the summation over all possible image embeddings I′I′, normalizing the probability.

Logarithm and Negative Sign: The negative log ensures that the loss decreases as the similarity of the correct text-image pair increases relative to other pairs.

This loss function trains the model to maximize the similarity of matching text-image pairs while minimizing the similarity of non-matching pairs.

Dual-Path Contrastive Loss: In practice, the contrastive loss is computed for both the Text-to-Image Alignment, which is Matching a text embedding TT with its correct image embedding II, and the Image-to-Text Alignment, matching an image embedding II with its corresponding text embedding TT.

The VQGAN architecture consists of an Encoder, which maps input images x to a latent representation ze(x). It has a Quantizer where Vector quantization is applied to ze(x), mapping it to the closest discrete vector zq in a learned codebook. Further, the Decoder Decodes zq back into the reconstructed image G(zq). Conversely, the Discriminator guides the generator via adversarial training, making reconstructed images visually realistic. The Vector Quantized Generative Adversarial Network (VQGAN) integrates adversarial training with vector quantization and a reconstruction objective to learn discrete latent representations.

The total loss function for the VQGAN is given by:

$$LVQGAN = LGAN + Lrecon + Lcommit \tag{2}$$

The Loss Components for VQGAN include:

Adversarial Loss (LGAN): This loss is observed from the adversarial training framework, where the generator aims to generate pragmatic images, and the discriminator differentiates between real and generated images. Mathematically, it is expressed as given below.

$$LGAN = E[logD(x)] + E[log(1 - D(G(z)))] \tag{3}$$

where:

D(x): Output of the discriminator for a real image xx.

G(z)G(z): Generated image from latent code z by the generator.

Reconstruction Loss (Lrecon): Encourages the generator to produce images that closely resemble the original inputs after encoding and decoding through the latent space. It is commonly implemented as an $\ell 1$\ell_1$\ell 1$ or $\ell 2$\ell_2$\ell 2$ norm:

$$Lrecon = \| x - G(E(x)) \| p \tag{4}$$

where:

E(x): Encoder output (latent representation).

p: Typically, 11 (MAE) or 22 (MSE).

Commitment Loss (Lcommit): Ensures that the Encoder efficiently uses the learned discrete codebook by penalizing large deviations between the encoder output and its nearest codebook vector.

$$Lcommit = \beta \| ze(x) - sg[zq] \| 22 \tag{5}$$

where:

ze(x): Latent representation output by the Encoder.

zq: Closest vector from the discrete codebook.

sg[·]: Stop gradient operation to prevent gradients from flowing into the codebook during optimization.

β: Hyperparameter controlling the weight of the commitment loss.

Fréchet Inception Distance (FID): The evaluation of the model was conducted using alternative methods as Fréchet Inception Distance (FID). FID is used in evaluating generative models, particularly in the context of image synthesis. FID is very popular because it quantifies both the quality and diversity of generated images in a single score and both quality and diversity are the most important evaluation measures of GANs. Other pixel-wise metrics such as Mean Squared Error (MSE) or Structural Similarity Index (SSIM), which are often too rigid to capture perceptual quality, FID compares the distributions of real and generated images in a feature space. The characteristic of FID is its alignment with human perception of the generated image quality. It captures high-level semantic features. It evaluates the entire distribution of generated images, making it sensitive to mode collapse. Inception Score is also used in measuring image quality but for our dataset there are biasing issues because ImageNet is a distinct and well-defined categories dataset. Fréchet Inception Distance (FID) is stated as shown in Eq. (6).

$$FID = \| \mu 1 - \mu 2 \| 2 + Tr(\Sigma 1 + \Sigma 2 - 2(\Sigma 1 \Sigma 2)1/2) \tag{6}$$

where μ1 and μ2 are the mean feature vectors of the generated and real images, and Σ1 and Σ2 are their covariance matrices.

Inception Score (IS) which is another evaluation parameter, can not be applied here because of the ImageNet dataset characteristics. However, we have shown best and worst results achieved through our models in Table 3. In addition to the results shown in Table 3, we have also compared our model results with other two SOTA approaches of image generation StackGAN++ and Dall-E and represented results in Table 4. Since this is a text quality dependent model, its limitations related to text to speech generation are also mentioned in the discussion section. Human inspection is also required to validate the GANs results so the generated images were validated by our team and some of the students who work in the similar domain. Our model has the potential of generating high quality realistic images in various categories and shows its robustness.

Table 3: FID score for generated images

Prompt	Generated image	Real image	FID score
Komodo dragon			31.250
Tennis ball			28.750
Tarantula			35.125
Teddy bear			40.375
Bear			42.76
A man			13.24

(Continued)

Table 3 (continued)

Prompt	Generated image	Real image	FID score
Shaktimaan-superhero			17.12

Table 4: Comparison of proposed model with existing models

Prompt	CLIP + VQGAN	StackGAN++	Dall-E
Shaktimaan-superman			
Teddy bear			
Bear			
Komodo dragon			
Tarantula			
Tennis ball			

(Continued)

Table 4 (continued)

Prompt	CLIP + VQGAN	StackGAN++	Dall-E
A Man			

The dataset used for this particular experiment is the ImageNet dataset, first introduced in 2009 [19]. This dataset is built on the WordNet structure and contains over 30 million annotated images, serving as benchmark standards for numerous computer vision applications. The implementation of VQGAN used by the authors of this paper learns from a codebook that encodes images as tokens. In this version, a reduction actor of 16 is utilized. Therefore, an image of size 256 × 256 would be normalized to 16 × 16 as shown in Fig. 4.

Figure 4: Sample images tested in this research [19]

CLIP's robust semantic understanding can mitigate minor ASR errors in speech-to-text conversion, ensuring the generated images remain relevant. VQGAN + CLIP can handle abstract or vague descriptions often found in spoken language. It also supports real-time applications like interactive art tools. Traditional GANs have limited semantic text-image matching but CLIP+VQGAN can easily generalize the match and produce high quality images. The highlight of our model is it works with unpaired data in the unsupervised learning environment unlike other traditional models. DALL-E is computationally intensive. It also has less control over style compared to other techniques like VQGAN + CLIP or diffusion models. Even stable diffusion models have limitations such as slow generation and also, they are resource intensive. Prompt sensitivity is also another problem with diffusion models. The objective of this paper is to take input speech from the user, convert it into text using the Chrome Web-to-Speech API, and then feed this text into our proposed VQGAN + CLIP model to generate beautiful imagery as shown in Fig. 3.

The authors aim to establish an optimal model for image generation using speech prompts and enhance the quality of the generated image. The text prompt provided to the model will include specific words and phrases that should be included in the resulting image. The model will encode these prompts and utilize a cosine similarity loss function to assess the similarity between the encoded text and the encoded image, ensuring they are indistinguishable. The model will keep on updating its latent space parameters until the desired image is generated. The model is divided into three parts: I. CLIP, II. GAN, III. Speech-to-text translation Contrastive Language Image Pretraining (CLIP) is a neural network model that has already been trained on over 400 million (image, text) pairs. As its name implies, the architecture produces a relevant image for the provided caption to the model. The rationale for selecting CLIP primarily lies in its ability to rapidly learn visual concepts through natural language supervision (Table 2). CLIP can be applied to any visual classification benchmark similar to the zero-shot capabilities of GPT-2 and GPT-3 by merely providing the names of the visual categories to be recognized. The CLIP model is built using the following sub-models, as shown in Figs. 5 and 6. Text Encoder and Image Encoder, these sub-models will embed the text and images to a mathematical space, and their dot product is used to check the similarity score. Fig. 5 shows the basic architecture and Fig. 6 shows the vision transformer.

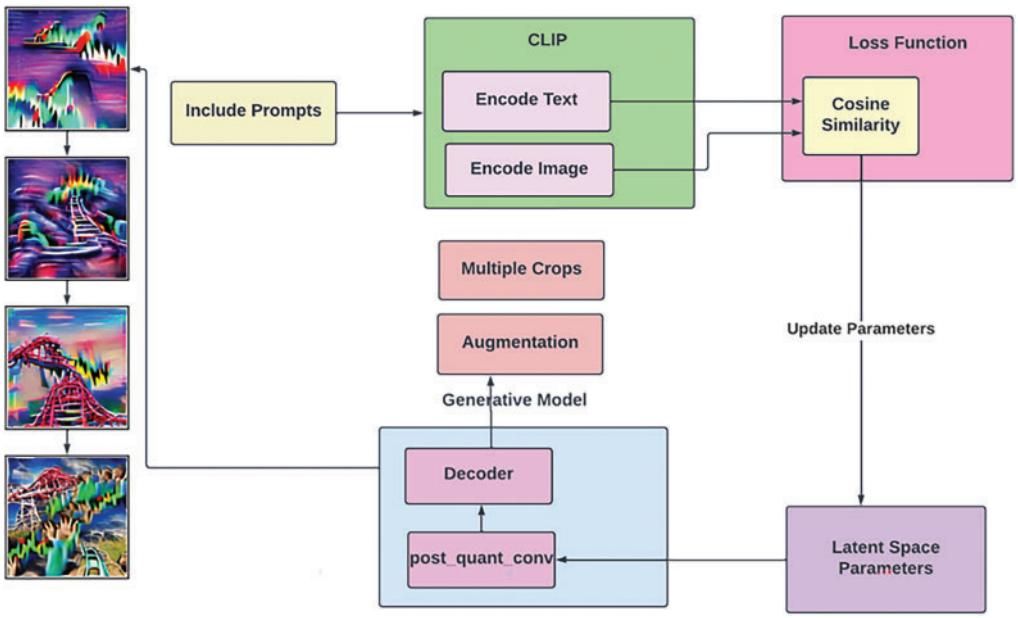

Figure 5: Block diagram of the proposed method

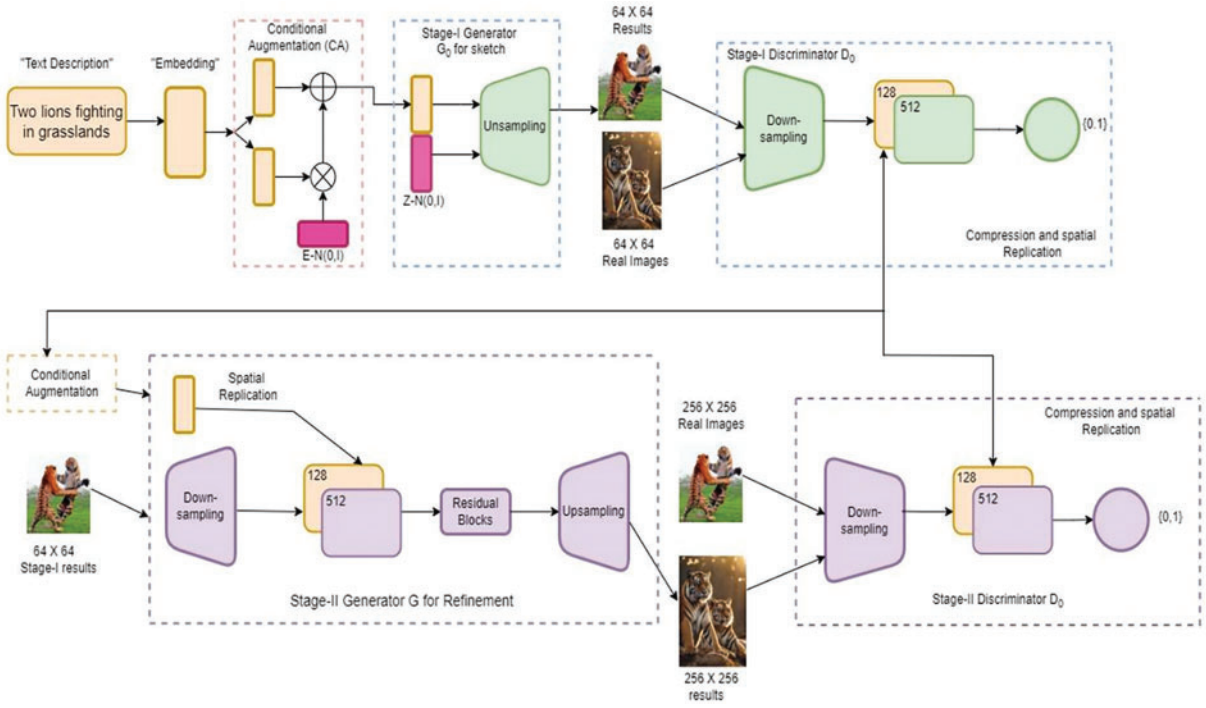

Figure 6: Basic architecture used in CLIP

The Speech-to-Text model used in this approach is the JavaScript Web Speech Application Programming Interface (API). Its ability to recognize language is affected by a lot of external factors such as speech clarity, browser support, speaker accent, and pronunciation. Variations in speech recognition accuracy can impact the final image generation, potentially leading to less accurate visual representations. To ensure optimal speech transcription, the model uses word processing techniques such as a language model and lexicon to optimally match the processed words to proper keywords to extract the most information from the transcription. This is done via the CLIP transformer as explained in the section proposed approach.

The accuracy of Speech-to-Text translation plays a pivotal role in the quality of the images generated by our model. Inaccuracies in the Speech-to-Text output, such as misinterpretations of unclear speech or background noise, can lead to incorrect or unintended visual representations. For instance, ambiguous or misrecognized words may result in images that do not align with the intended speech input. To mitigate these issues, this model can be integrated with more robust Speech-to-Text systems that are better at handling diverse speech inputs, including accents and noisy environments. Future work may also explore preprocessing techniques to refine the Speech-to-Text output before feeding it into the image generation pipeline, thereby enhancing the system's overall reliability.

6 Results and Discussion

The images shown in Table 2, display the results obtained by the Generative Adversarial Networks on the ImageNet dataset. It was noted that the quality of the generated images improves with an increase in the number of iterations. However, it was also noted that an increase in the number of iterations also correlates with an increase in the computational time. The authors explored the effects of iteration count on image quality and computational cost. The images generated from the textual description of "Roller-coaster" over various iterations have been presented below. The images produced using the proposed method also contain

some amount of noise and, thus, are not very sharp due to the large number of classes in the dataset and the various weights that were assigned to each class by the CLIP transformer. Nevertheless, they still incorporate elements from the provided input prompts, illustrating that the model is capable of capturing data from verbal input and generating a corresponding synthesized image despite a low iteration count.

The authors calculated the FID score for the outputs as a quantitative measure to assess the model [50]. The FID score is a popular statistic for evaluating the quality of produced pictures compared to actual photos, measuring how closely the distribution of characteristics in produced pictures resembles that of real photos. However, the FID score is not sufficient to represent the characteristics of a model designed for any newly generated content.

The Inception Score (IS) and FID are very popular evaluation metrics for checking the quality of generated images [51]. The IS shows how well the generated images resemble real images by assessing their diversity. Higher IS values indicate more diverse and realistic images. The FID score, compares the distribution of features extracted from real and generated images, with lower scores indicating closer similarity to real images. These metrics were chosen for their effectiveness in capturing both the quality and diversity of generated images, making them ideal for evaluating the model's performance. It is calculated by first using a pre-trained InceptionV3 network to extract feature vectors from both sets of images. The mean and covariance of these feature vectors are then computed for both the real images and the generated images. The FID score is determined by the Frechet distance formula, which measures the distance between the two multivariate Gaussian distributions represented by these means and covariances.

The model achieved an Inception Score of 72 and an FID score of 28.75, showcasing its capability to produce high-quality and diverse images. In this study, the FID score was relatively low, mainly because the VQGAN f16 artifacts were penalized harshly. Despite the noisy generated images, the authors emphasize that the proposed model's intended purpose is to generate abstract visuals using speech inputs. The outputs of prompts containing descriptions of objects do not present in the ImageNet dataset used by VQGAN, or the zero-shot generated images, demonstrate the diverse nature of the model and its outputs. The model can generate unique, highly imaginative, and abstract images that inspire artistic creativity.

The authors acknowledge that the proposed method has limitations, and the images generated are not perfect due to the large number of classes in the dataset and the various weights assigned to each class by the CLIP transformer. When generating an image from a textual prompt, the GANs and transformers attempt to map the textual input to an image output. However, the mapping is not always straightforward, especially when dealing with abstract concepts or unusual descriptions. As a result, the model sometimes produces noisy and blurry images. However, they believe that their research provides a promising proof-of-concept for future work on the use of speech prompts and GANs for generating imaginative and abstract images. Additionally, the authors suggest extending the method to generate animations and videos using speech prompts, opening up new possibilities for creative expression. The diversity of the dataset introduces some level of noise into the model's understanding of the images, which can lead to the generation of suboptimal images. Despite these limitations, the authors believe that their method demonstrates promising results and opens up new possibilities for creative expression.

Handling the computational demands associated with fine-tuning the CLIP and VQGAN models is challenging. Due to limited processing capacity, it was crucial to optimize the training process to ensure efficient use of resources. Employed techniques such as model pruning and mixed-precision training to reduce the computational load while maintaining the quality of the generated images. Additionally, the use of noise reduction and normalization techniques during preprocessing to handle variations in speech input, such as different accents, speech speeds, and background noises, significantly affects the accuracy of the Speech-to-Text translation.

The generated images contain noise and are not very sharp due to the large number of classes in the dataset and the various weights assigned to each class by the CLIP transformer; they contain elements from the provided input prompts. Therefore, the model can capture information through speech input and generate a coinciding synthesized image despite the low iteration count. However, the intended purpose of the proposed model is to generate abstract visuals using speech inputs, and the outputs of prompts containing descriptions of objects not present in the ImageNet dataset used by the VQGAN exhibit the diverse nature of the model and its outputs shown in Fig. 7.

Figure 7: Roller-coaster generated by CLIP

The authors have also conducted a quantitative analysis of the model by calculating the FID score for the outputs. However, they believe that a model intended for vague and imaginative image generation can have numerous use cases, and the FID score does not fully capture these capabilities. The FID score was obtained using the open-source implementation of FID Score—"pytorch-FID" as shown in Table 2.

The "FID" measures the similarity between two images. It is evident that the FID score is relatively low, largely due to the VQGAN f16 artifacts being heavily penalized as highlighted in Fig. 8. Table 3 displays the comparison of outputs of the suggested model with StackGAN++ and DALL-E. The observed output exhibits the CLIP+VQGAN's capability to capture the exact images. It can be observed the closeness of the generated image according to prompt and learned images from the dataset provided. The ability of the CLIP + VQGAN architecture to interpret and generate images based on speech input depends heavily on how well it processes and aligns the meaning of the input with visual representations. Abstract prompts (e.g., "freedom," "a dream of stars") and concrete prompts (e.g., "a red car on a sunny road") present distinct challenges and opportunities for evaluation. As the model is trained on a limited dataset, which does not include diverse

ranges of accents, this also degrades the model's performance. The data primarily consists of standard or regional accents, which may require the system to accurately recognize words spoken in different accents. This inhibition puts a limit on the performance of the model. However, accents can alter the pronunciation of words significantly. Different accents may emphasize or de-emphasize specific phonemes (sounds), leading to misinterpretation by the recognition algorithms. Different speech patterns, such as speed or clarity, can sometimes accompany accents. Another challenge is the systems can adapt to individual speakers over time. Still, this adaptation may only be effective for some accents, particularly if the system has yet to be trained in similar voices. Table 3 also showcases the borderline cases. The comparison of the model with standard techniques is shown in Table 4. This gives the validation of the output generated by the suggested model. However, ImageNet is a comprehensive, labeled visual dataset that has been crucial in advancing computer vision research. Its vast collection of images across 1000 object categories has become a standard benchmark for training and evaluating models, particularly in tasks such as image classification and object detection.

Figure 8: Images generated by CLIP

The suggested model produced decent results using less computational resources. Though the FID score is higher it could produce real-time images. FID is more due to poor image quality, mode collapse, inadequate training, feature mismatches, or a lack of diversity in either the generated images or the real dataset. Speech/audio related challenges also increase FID score. As the model is trained on a limited dataset, which does not include diverse ranges of accents, this also degrades the model's performance. The data primarily consists of standard or regional accents, which may require the system to accurately recognize words spoken in different accents. This inhibition puts a limit on the performance of the model. However, accents can alter the pronunciation of words significantly. Different accents may emphasize or de-emphasize specific phonemes (sounds), leading to misinterpretation by the recognition algorithms. Different speech patterns, such as speed or clarity, can sometimes accompany accents. Another challenge is the systems

can adapt to individual speakers over time. Still, this adaptation may only be effective for some accents, particularly if the system has yet to be trained on similar voices.

The following are the limitations of this study:

- CLIP often excels at identifying familiar things, but it suffers from more complex or systematic tasks. For highly fine-grained categorization, such as determining the differences between different automobile models, zero-shot CLIP performs poorly compared to task-specific models.
- Additionally, CLIP still struggles to generalize to photos that were not included in its pretraining dataset.
- CLIP's zero-shot classifiers can be delicate to phrasing and may require "prompt engineering" through trial and error to function successfully.
- The speech prompts are translated to text using the Chrome Web-to-Speech API. This may not be fully accurate and can result in errors when translating words with various accents. The authors plan to follow up on this work by working on speech encoding prompts for image translation in the future.
- The proposed model encounters challenges with ambiguous or unclear speech inputs, often resulting in less accurate or relevant image generation. For highly abstract concepts, the model may struggle to produce coherent visual outputs due to the complexity of translating abstract speech into visual elements.
- One of the major challenges in this implementation was handling the computational demands associated with fine-tuning the CLIP and VQGAN models. Due to limited processing capacity, it was crucial to optimize the training process to ensure efficient use of resources. The employed techniques such as model pruning and mixed-precision training to reduce the computational load while maintaining the quality of the generated images.
- Additionally, there are ethical considerations, such as the potential for misuse in generating misleading or inappropriate content. It is essential to implement guidelines and safeguards to ensure ethical use and address these challenges. Future research should focus on improving the model's robustness and developing ethical frameworks to guide its application.

7 Conclusion

This research introduced a framework for synthesizing images based on speech inputs, approached through three stages. Firstly, the model processes the voice input and extracts meaningful words using the speech module available in web browsers such as Chrome, Edge, and Firefox. Subsequently, the phrase is tokenized, and the model generates a text string with essential characteristics for the CLIP transformer to use as input. The image encoder utilizes zero-shot prediction to convert these tokens into a vector that the VQGAN can employ to produce image cropping. This technology holds promise for generating quasi-realistic graphic artistry and provides utility for the gaming industry in developing interactive components. The primary motivation for this research has been its potential applications in education, entertainment, and visual art fields.

Specifically, within the field of education, the model could be used to create visual aids for language learning, where students can generate images based on spoken descriptions, aiding in vocabulary acquisition and comprehension. In the field of art, artists could use the model to visualize their spoken ideas, fostering creative expression and exploration. In design, the model could assist designers in quickly prototyping visual concepts based on verbal descriptions, streamlining the design process. For instance, a case study in an art workshop demonstrated that participants were able to generate unique visual pieces by describing their concepts verbally, which the model then transformed into visual art. This will be a paradigm shift in these fields and would radically expand the boundaries of interface for artists, designers, and educators. The authors aim to enhance this study by focusing on audio-derived speech encoding and exploring alternative speech embedding techniques that could refine a bipartite model. The work can also be extended using

stable diffusion models in different applications. Future research could focus on enhancing the model's ability to interpret and generate images from abstract or ambiguous prompts, possibly through advanced natural language processing techniques and more extensive training datasets.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney. Moreover, supported by the Researchers Supporting Project, King Saud University, Riyadh, Saudi Arabia, under Ongoing Research Funding (ORF-2025-14).

Author Contributions: Conceptualization, Smita Mahajan, Shilpa Gite; methodology, Shaunak Inamdar, Deva Shriyansh, Akshat Ashish Shah, Shruti Agarwal, Shilpa Gite, Smita Mahajan; software, Shilpa Gite, Smita Mahajan; validation, Shaunak Inamdar, Deva Shriyansh, Akshat Ashish Shah, Shruti Agarwal, Shilpa Gite, Smita Mahajan, Biswajeet Pradhan, Abdullah Alamri; formal analysis, Shaunak Inamdar, Deva Shriyansh, Akshat Ashish Shah, Shruti Agarwal; investigation, Shaunak Inamdar, Deva Shriyansh, Akshat Ashish Shah, Shruti Agarwal, Shilpa Gite, Smita Mahajan, Biswajeet Pradhan, Abdullah Alamri; resources, Biswajeet Pradhan; data curation, Shaunak Inamdar, Deva Shriyansh, Akshat Ashish Shah, Shruti Agarwal; writing—original draft preparation, Shaunak Inamdar, Deva Shriyansh, Akshat Ashish Shah, Shruti Agarwal; writing—review and editing, Shilpa Gite, Smita Mahajan, Biswajeet Pradhan, Abdullah Alamri; visualization, Shilpa Gite, Smita Mahajan, Biswajeet Pradhan, Abdullah Alamri; supervision, Shilpa Gite, Smita Mahajan; project administration, Shilpa Gite, Smita Mahajan, Biswajeet Pradhan; funding acquisition, Biswajeet Pradhan, Abdullah Alamri. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: ImageNet Dataset is having Free Access for Research. It is primarily intended for non-commercial academic research. Researchers need to create an account and request access to download the dataset. The link to the version of the code is available at the link provided here https://colab.research.google.com/drive/1MJeP6z4opQC6sA9b20kjkh_HGXzFIyyV?usp= sharing (accessed on 25 January 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Siarohin A, Sangineto E, Lathuilière S, Sebe N. Deformable GANs for pose-based human image generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 3408–16. doi:10.1109/CVPR.2018.00359.
2. Baraheem SS, Le TN, Nguyen TV. Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook. Artif Intell Rev. 2023;56(10):10813–65. doi:10.1007/s10462-023-10434-2.
3. Liz-López H, Keita M, Taleb-Ahmed A, Hadid A, Huertas-Tato J, Camacho D. Generation and detection of manipulated multimodal audiovisual content: advances, trends and open challenges. Inf Fusion. 2024;103(2):102103. doi:10.1016/j.inffus.2023.102103.
4. Arjovsky M, Chintala S, Bottou L, Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia: ACM; 2017. p. 214–23. doi:10.5555/3305381.3305404.
5. Andonian A, Osmany S, Cui A, Park Y, Jahanian A, Torralba A, et al. Paint by word. arXiv:2103.10951. 2021.
6. Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, et al. StackGAN: realistic image synthesis with stacked generative adversarial networks. IEEE Trans Pattern Anal Mach Intell. 2019;41(8):1947–62. doi:10.1109/TPAMI.2018.2856256.
7. Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. arXiv:1809.11096. 2018.

8. Lin ZY, Geng S, Zhang R, Gao P, De Melo G, Wang X, et al. Frozen CLIP models are efficient video learners. In: Computer Vision—ECCV 2022: 17th European Conference; 2022 Oct 23–27; Tel Aviv, Israel. 2022. p. 388–404. doi:10.1007/978-3-031-19833-5_23.
9. Li J, Tang T, Zhao WX, Nie JY, Wen JR. Pre-trained language models for text generation: a survey. ACM Comput Surv. 2024;56(9):1–39. doi:10.1145/3649449.
10. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inform Process Syst. 2020;33:1877–901.
11. Chen M, Radford A, Child R, Wu J, Jun H, Luan D, et al. Generative pretraining from pixels. In: Proceedings of the 37th International Conference on Machine Learning; 2020; Vienna, Austria. p. 1691–703.
12. Dolhansky B, Ferrer CC. Eye in-painting with exemplar generative adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 7902–11. doi:10.1109/CVPR.2018.00824.
13. Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA: IEEE; 2021. p. 12868–78. doi:10.1109/cvpr46437.2021.01268.
14. Frolov S, Hinz T, Raue F, Hees J, Dengel A. Adversarial text-to-image synthesis: a review. Neural Netw. 2021;144(4):187–209. doi:10.1016/j.neunet.2021.07.019.
15. Ge S, Hayes T, Yang H, Yin X, Pang G, Jacobs D, et al. Long video generation with time-agnostic VQGAN and time-sensitive transformer. In: Computer Vision–ECCV 2022: 17th European Conference; 2022 Oct 23–27; Tel Aviv, Israel. 2022. p. 102–18.
16. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. arXiv:1406.2661. 2014.
17. Ling H, Kreis K, Li D, Kim SW, Torralba A, Fidler S. EditGAN: high-precision semantic image editing. Adv Neural Inform Process Syst. 2021;34:16331–45.
18. Donahue J, Dieleman S, Bińkowski M, Elsen E, Simonyan K. End-to-end adversarial text-to-speech. arXiv:2006.03575. 2020.
19. Deng J, Dong W, Socher R, Li LJ, Kai L, Li FF. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA: IEEE; 2009. p. 248–55. doi:10.1109/CVPR.2009.5206848.
20. Park K, Mott BW, Min W, Boyer KE, Wiebe EN, Lester JC. Generating educational game levels with multistep deep convolutional generative adversarial networks. In: IEEE Conference on Games (CoG); 2019 Aug 20–23; London, UK: IEEE; 2019. p. 1–8. doi:10.1109/cig.2019.8848085.
21. Crowson K, Biderman S, Kornis D, Stander D, Hallahan E, Castricato L, et al. VQGAN-CLIP: open domain image generation and editing with natural language guidance. In: Computer Vision–ECCV 2022; 2022; Cham, Switzerland: Springer Nature Switzerland. p. 88–105. doi:10.1007/978-3-031-19836-6_6.
22. Orynbay L, Razakhova B, Peer P, Meden B, Emeršič Ž. Recent advances in synthesis and interaction of speech, text, and vision. Electronics. 2024;13(9):1726. doi:10.3390/electronics13091726.
23. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 8107–16. doi:10.1109/cvpr42600.2020.00813.
24. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. arXiv:1710.10196. 2017.
25. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE; 2019. p. 4401–10. doi:10.1109/cvpr.2019.00453.
26. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: a survey. ACM Comput Surv. 2022;54(10s):1–41. doi:10.1145/3505244.
27. Li XL, Liang P. Prefix-tuning: optimizing continuous prompts for generation. arXiv:2101.00190. 2021.

28. Li X, Zhang J, Liu Y. Speech driven facial animation generation based on GAN. Displays. 2022;74:102260. doi:10.1016/j.displa.2022.102260.
29. Meyer S, Tilli P, Denisov P, Lux F, Koch J, Vu NT. Anonymizing speech with generative adversarial networks to preserve speaker privacy. In: 2022 IEEE Spoken Language Technology Workshop (SLT); 2023 Jan 9–12; Doha, Qatar: IEEE; 2023. p. 912–9. doi:10.1109/SLT54892.2023.10022601.
30. Liu MY, Tuzel O. Coupled generative adversarial networks. arXiv:1606.07536. 2016.
31. Mansimov E, Parisotto E, Ba JL, Salakhutdinov R. Generating images from captions with attention. arXiv:1511.02793. 2015.
32. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv:1411.1784. 2014.
33. Mokady R, Hertz A, Bermano AH. ClipCap: clip prefix for image captioning. arXiv:2111.09734. 2021.
34. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML); 2021. Vol. 139, p. 8748–63.
35. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, et al. Photorealistic text-to-image diffusion models with deep language understanding. Adv Neural Inf Process Syst. 2022;35:36479–94.
36. Li H, Xu F, Lin Z. ET-DM: text to image via diffusion model with efficient Transformer. Displays. 2023;80(1):102568. doi:10.1016/j.displa.2023.102568.
37. Chen J, Yu J, Ge C, Yao L, Xie E, Wu Y, et al. PixArt-α: fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv: 2310.00426. 2023.
38. Seneviratne S, Senanayake D, Rasnayaka S, Vidanaarachchi R, Thompson J. DALLE-URBAN: capturing the urban design expertise of large text to image transformers. In: 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA); 2022 Nov 30–Dec 2; Sydney, Australia: IEEE; 2022. p. 1–9. doi:10.1109/DICTA56598.2022.10034603.
39. Radford A, Metz L, Chintala S, Dinakaran R, Easom P, Zhang L, et al. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434. 2015.
40. Patashnik O, Wu Z, Shechtman E, Cohen-Or D, Lischinski D. StyleCLIP: text-driven manipulation of StyleGAN imagery. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada: IEEE; 2021. p. 2085–94. doi:10.1109/ICCV48922.2021.00209.
41. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-shot text-to-image generation. In: Proceedings of the International Conference on Machine Learning (ICML); 2021. Vol. 139, p. 8821–31.
42. Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H. Generative adversarial text-to-image synthesis. In: Proceedings of the International Conference on Machine Learning (ICML); 2016. Vol. 49, p. 1060–9.
43. Song H, Dong L, Zhang WN, Liu T, Wei F. CLIP models are few-shot learners: empirical studies on VQA and visual entailment. arXiv:2203.07190. 2022.
44. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS 2017); 2017; Curran Associates, Inc. Vol. 30.
45. Tulyakov S, Liu MY, Yang X, Kautz J. MoCoGAN: decomposing motion and content for video generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 1526–35. doi:10.1109/CVPR.2018.00165.
46. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6. doi:10.1038/323533a0.
47. Tian Y, Ren J, Chai M, Olszewski K, Peng X, Metaxas DN, et al. A good image generator is what you need for high-resolution video synthesis. arXiv:2104.15069. 2021.
48. Wang X, Qiao T, Zhu J, Hanjalic A, Scharenborg O. S2IGAN: speech-to-image generation via adversarial learning. arXiv:2005.06968. 2020.
49. Wu H, Zheng S, Zhang J, Huang K. GP-GAN: towards realistic high-resolution image blending. In: Proceedings of the 27th ACM International Conference on Multimedia; 2019; Nice France: ACM. p. 2487–95. doi:10.1145/3343031.3350944.

50. Wu J, Zhang C, Xue T, Freeman WT, Tenenbaum JB. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In: Advances in Neural Information Processing Systems (NeurIPS 2016); 2016; Curran Associates, Inc. Vol. 30.
51. Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 1316–24. doi:10.1109/CVPR.2018.00143.
52. Zhu M, Pan P, Chen W, Yang Y. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE; 2019. p. 5802–10. doi:10.1109/CVPR.2019.00595.
53. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy: IEEE; 2017. p. 2223–32. doi:10.1109/ICCV.2017.244.
54. Borji A. Pros and cons of GAN evaluation measures: new developments. Comput Vis Image Underst. 2022;215(4):103329. doi:10.1016/j.cviu.2021.103329.
55. Figueira A, Vaz B. Survey on synthetic data generation, evaluation methods and GANs. Mathematics. 2022;10(15):2733. doi:10.3390/math10152733.