International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-June-2026, ISSN (Online): 2395-566X

Text to Image Generation Using Gen AI

Teju M, Assistant Professor Divakar K M
Department of Computer Science and Engineering, SJC Institute of Technology, Chikkaballapur, Karnataka, India

Abstract — Generative Artificial Intelligence has changed the way we make images from text. Now we can make quality images from what we write. This is because of models that use special architectures. These models are really good at making images that look real. Are about the right thing. We can use these models to make art, design and ads. They are also useful in education, healthcare and gaming. This saves us time and money because we do not have to make images by hand. This paper is about how we can make images from text using Generative Artificial Intelligence. We look at how the models work and what's new about them. We talk about models like Stable Diffusion, DALL·E and Imagen. We look at how the whole process works, from getting the text ready to making the image. We also think about how to make the images better by using the words. We discuss what is good and bad about the models we have now. We also look at what other people have found out about making images from text. We compare the ways to do it and talk about what is new and interesting. We think about how we can make images that're just what we want and how we can make the models work better and faster. Generative Artificial Intelligence models that use diffusion are good at making images that look real. Are about the right thing. They open up possibilities, for art and industry. This paper ends by talking about what we need to do to make the models better and more responsible.

Keywords— Text-to-Image Generation, Generative AI (Gen AI), Artificial Intelligence, Diffusion Models, Prompt Engineering, Image Synthesis, AI Image Generator

I. INTRODUCTION

Artificial Intelligence (AI) has changed quickly in recent years, with Generative Artificial Intelligence (GenAI) becoming one of its most significant branches. Unlike traditional AI systems that mainly analyze or classify data, Generative AI can create new content, such as text, images, audio, video, and code. Among these applications, text-to-image generation has received a lot of attention because it can turn natural language descriptions into realistic and high-quality images. This technology allows users to make visual content just by providing text prompts, making image creation more accessible and efficient in various fields.

Recent advances in deep learning, especially transformer-based language models and diffusion models, have greatly improved the quality and realism of generated images. Top models like DALL·E, Stable Diffusion, Imagen, and FLUX have shown impressive performance in understanding complex text descriptions and producing visually coherent images. These models use large datasets and multimodal learning methods to create strong links between text and visual information, allowing them to generate images that closely match user intent.

Text-to-image generation has many practical uses across industries. In graphic design and digital art, it speeds up the creative process by generating concept artwork and illustrations. In advertising and marketing, it helps quickly create promotional materials and product visualizations. Technology is also valuable in education for creating instructional content, in healthcare for medical visualizations, in architecture for conceptual building designs, in entertainment and gaming for creating characters and environments, and in e-commerce for product visualization. These applications show the growing importance of Generative AI in boosting productivity, creativity, and decision-making.

Despite these advances, several challenges persist. The quality of generated images relies heavily on how clear and specific the input prompt is. Models might produce inaccurate or biased outputs due to limitations in training data or the complexity of natural language. Furthermore, issues related to copyright, misinformation, ethical use, computing costs, and responsible deployment continue to draw significant attention from researchers and policymakers. It's important to address these challenges to ensure the reliable and ethical use of text-to-image generation systems.

This paper presents a thorough study of text-to-image generation using Generative AI. It reviews the development of image generation techniques, explains how modern diffusion-based architecture works, compares leading text-to-image models, and discusses their advantages, limitations, and real-world applications. The paper also points out current research trends and future directions aimed at improving image quality, computational efficiency, control, fairness, and responsible AI practices. The goal is to provide a clear understanding of the

technologies behind modern text-to-image generation and their potential impact on future intelligent content creation systems.

II. RELATED WORK

Text-to-image generation has changed quickly thanks to progress in deep learning, multimodal learning, and Generative Artificial Intelligence (GenAI). Early methods relied mainly on Generative Adversarial Networks (GANs), which could create realistic images from random noise. Reed et al. (2016) were among the first to introduce a text-conditioned GAN model, proving that textual descriptions could effectively guide image creation. While GAN-based methods produced visually appealing images, they often faced issues like training instability, mode collapse, and weak connections between the text and the generated images.

To address these problems, Attain (Xu et al., 2018) added an attention mechanism. This allowed the model to focus on specific words in the input text while generating different parts of an image. This change greatly improved image quality and semantic consistency. Following this, DM-GAN (Zhu et al., 2019) included a dynamic memory module to enhance the generated images and better capture the fine details described in the text.

The introduction of transformer-based vision-language models further pushed forward text-to-image generation. OpenAI's CLIP (Contrastive Language-Image Pre-training) learned to create joint representations of images and text by training on large image-caption datasets. CLIP improved the connection between text and images and became a crucial part of many modern image generation frameworks.

A significant breakthrough came with diffusion models. Unlike GANs, diffusion models create images by gradually removing noise from random inputs through multiple steps. This method leads to higher image quality and greater stability during training. GLIDE (Nichol et al., 2021) showcased the capability of diffusion models to produce photorealistic images from text descriptions while allowing for image editing.

Then, DALL·E 2 (Ramesh et al., 2022) combined CLIP embeddings with diffusion-based image generation to create highly realistic and accurate images. Around the same time, Imagen (Sahara et al., 2022) showed that larger language models greatly improved understanding of prompts, resulting in better image quality. Stable Diffusion (Rombach et al., 2022) presented latent diffusion techniques that significantly lowered computing needs while keeping high-quality image creation, making it feasible for large-scale use on consumer hardware and speeding up research in open-source Generative AI.

Recent research has concentrated on improving controllability, personalization, and efficiency. ControlNet (Zhang et al., 2023) allowed users to influence image generation using extra conditions like sketches, edge maps, pose information, or depth maps while keeping prompt fidelity. Low-Rank Adaptation (LoRA) and DreamBooth have made model fine-tuning more effective, enabling customization with smaller datasets and lower computational expense. New models, such as SDXL and FLUX, have further improved image realism, adherence to prompts, and speed of generation through better architectures and training methods.

Despite the impressive progress, some research problems persist. Current models need substantial computational resources for both training and use, and the generated outputs may reflect biases from the training datasets.

Additionally, challenges related to copyright, intellectual property, misinformation, fairness, and ethical use continue to hinder widespread adoption. Thus, recent research focuses on responsible AI, effective model optimization, explainable image generation, and improved understanding of prompts to create more reliable and trustworthy text-to-image generation systems.

The reviewed literature shows that diffusion-based Generative AI models have become the leading method for text-to-image synthesis due to their high image quality, semantic accuracy, and scalability. However, more research is needed to improve computational efficiency, reduce bias, enhance interpretability, and promote responsible use in real-world applications.

III. RESEARCH METHODOLOGY

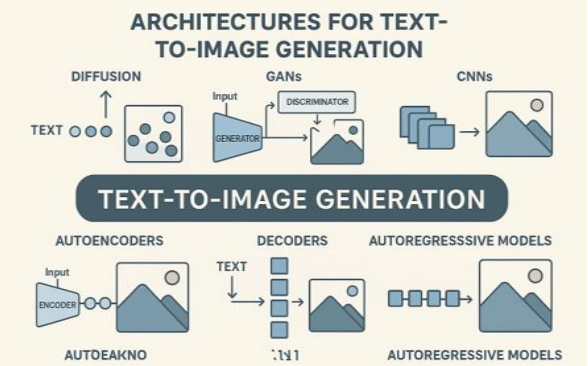


Figure 1: System Architecture of Text to Image Generation Using Gen AI

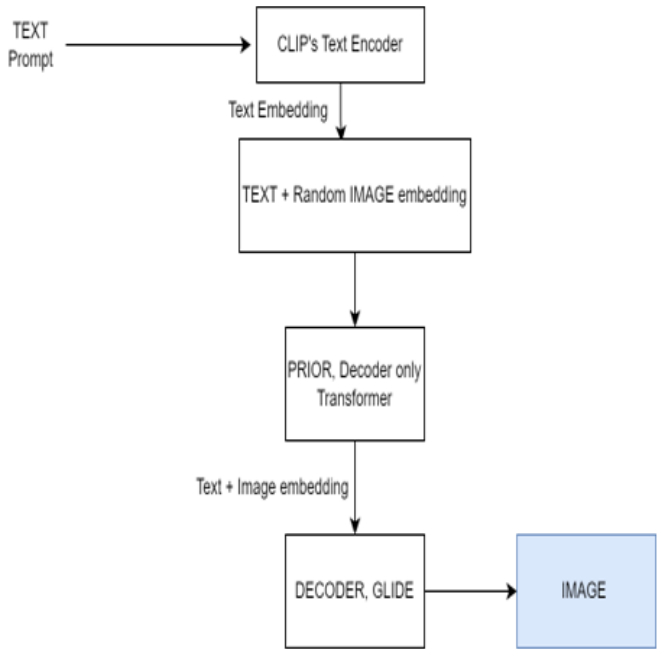
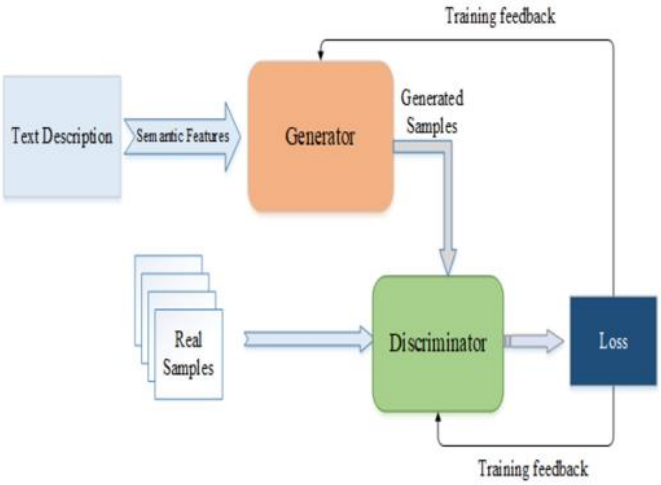


Figure 2: Flow Chart



Activity Diagram

1. Data Collection and Preprocessing

The dataset we use for text-to-image generation has paired text descriptions and images. We collect these from available datasets. Before we train the Generative AI model, we do some preprocessing to make the image generation better. Here are the steps we take:

- We get image-caption datasets from sources.
- We remove any duplicate or corrupted images.
- We. Standardize the text descriptions.
- We resize images to a fixed size.
- We normalize the pixel values of the images.
- We use a text encoder to turn text prompts into tokens.
- We split the dataset into training, validation and testing datasets.

These steps help our Generative AI model perform better and create high-quality images from text.

2. Generative AI Models

Stable Diffusion

Stable Diffusion is a type of Generative AI model that creates images from text prompts. It works by removing noise from a hidden representation to create realistic images.

DALL·E

DALL·E is another model that converts text into images. It uses a transformer to create creative images from natural language descriptions. DALL·E can generate images of different objects and scenes.

Imagen

Imagen is a text-to-image model developed by Google. It uses language understanding and diffusion models to create images that match the input text. Imagen generates images.

Midjourney

Midjourney is an AI model that creates images from text prompts. It is often used for art and design. Midjourney produces impressive images.

Generative Adversarial Networks (GANs)

GANs have two networks: a Generator and a Discriminator. The Generator creates images and the Discriminator checks if they are real. Through training GANs learn to generate images.

3. Performance Evaluation

We evaluate the performance of text-to-image generation models using these metrics:

- Fréchet Inception Distance (FID).
- Inception Score (IS).
- CLIP Score.
- Human Evaluation.
- Image Diversity.

These metrics help us assess the quality, realism, diversity and accuracy of the generated images. A lower FID and higher Inception Score and CLIP Score mean the model is better, at generating images that match the text prompts.

Research Gap

Most of the time people are trying to make one good picture from what someone writes. Not many people are working on making systems where people can talk to the computer and say what they like and do not like about the picture. Then the computer makes a new picture that is similar to the old one but also a little better. This way the computer can make pictures that are all related to each other. The goal is to make a system where people can have a conversation with the computer, about the pictures and the computer can make pictures based on what the person says and all the pictures look like they go together.

G Research Motivation

The Generative AI field is moving forward fast and this has made text-to-image models a lot better at creating realistic and high quality images from what people write.. Most systems only make one image from what you write and you cannot do much with it after it is made. If the image is not what you wanted you have to write everything which is a waste of time. This research is about making a text-to-image system that's more interactive and easier to use. Of just making one image we want to let people tell us what they think and then we can make the image better. We want people to be able to give us feedback in a way so we can keep making the image better and better and make sure it is still consistent with what we made before. This way we can understand what people like make sure the words and images match and create a way to make images that's just for each person.

The reason we are doing this research is to fill the gap between making one image and being able to have a conversation about the image. We want to make Generative AI systems more flexible and able to adapt so they can be used in the world for things like making digital art creating content, teaching, advertising and designing products. We think that research motivation like this can help make Generative AI better and more useful, for people. The main goal of this research motivation is to improve Generative AI and make it more user friendly

Research Motivation

The Generative AI field is moving forward fast and this has made text-to-image models a lot better at creating realistic and high quality images from what people write.. Most systems only make one image from what you write and you cannot do much with it after it is made. If the image is not what you wanted you have to write everything which is a waste of time. This research is about making a text-to-image system that's more interactive and easier to use. Of just making one image we want to let people tell us what they think and then we can make the image better. We want people to be able to give us feedback in a way so we can keep making the image better and better and make sure it is still consistent with what we made before. This way we can understand what people like make sure the words and images match and create a way to make images that's just for each person.

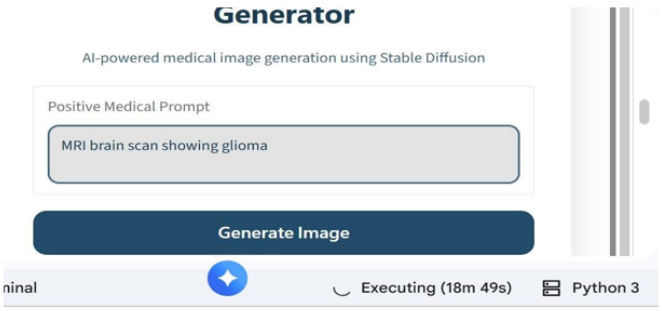
The reason we are doing this research is to fill the gap between making one image and being able to have a conversation about the image. We want to make Generative AI systems more flexible and able to adapt so they can be used in the world for things like making digital art creating content, teaching, advertising and designing products. We think that research motivation like this can help make Generative AI better and more useful, for people. The main goal of this research motivation is to improve Generative AI and make it more user friendly

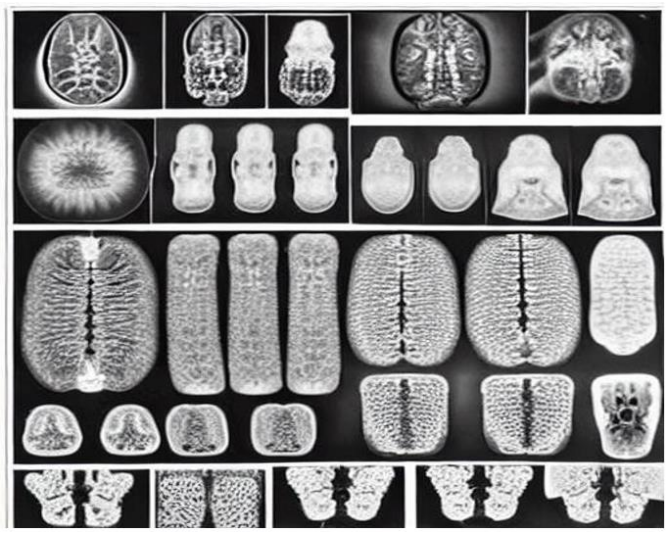
IV. RESULTS AND DISCUSSION

The Medical Text-to-Image Generation system that we made works well. It uses a type of Artificial Intelligence called Stable Diffusion to make pictures from text.

This system takes a sentence about medicine from the user. Makes a picture that shows what the sentence is talking about. We made a website where people can type in what they want to see like the name of a disease or what a radiology picture looks like. Then the Stable Diffusion model takes what they typed. Makes a picture that looks real.

We tried the system with the sentence "MRI brain scan showing glioma". The Medical Text-to-Image Generation system made a picture of a brain scan that shows what glioma looks like. The Medical Text-, to-Image Generation system did a job making the picture.





The Generative AI model that we made can make good medical pictures from what people write. These pictures look a lot like the pictures you get from MRI brain scans. They show what the person meant when they wrote the description.

The Stable Diffusion model is good at figuring out how what people write is connected to what you see in pictures. This means it can make pictures that look real without someone having to design them by hand.

Our system is useful, for reasons:

- It can make pictures straight from what people write.
- It is easy to use and understand.
- It makes pictures that look like they came from an MRI machine.
- It saves time when making pictures to help explain things.
- It can help people learn about medicine and do research by making examples that you can see.

V. CONCLUSION

This study is about a system that uses Generative AI to make pictures from words. It uses something called the Stable Diffusion model to do this. The system can take what doctors write. Turn it into pictures that look like they were made with an MRI machine. You can use it on a website. It is really easy to use. This shows that Generative AI can be very useful for making pictures.

The people who made this system tried it out. Found that it makes pictures that really look like what the doctors wrote. The pictures look real. They match what the words say. This means that the Stable Diffusion model can help make medical pictures for teachers, researchers and doctors to use. It saves time because you do not have to make the pictures by hand. The system is a way to make medical pictures from words and it is fast and easy to use. The Generative AI system is good, for making pictures for many different purposes like teaching and research and the Stable Diffusion model is a big part of it..

Acknowledgment

The authors want to say thank you to Mr. Divakar K. M., who's the Project Guide and Assistant Professor at the Department of Computer Science and Engineering at SJC Institute of Technology in Chikkaballapur, Karnataka. They are thankful for his help and support throughout the research work. Mr. Divakar K. M. Gave the authors advice and encouraged them to keep going.

The authors also want to thank the Department of Computer Science and Engineering at SJC Institute of Technology in Chikkaballapur, Karnataka. They are thankful for the facilities and equipment they got to use. The department gave them a place to work and do research. This helped the authors finish their project.

The authors are also thankful to the people who made the Stable Diffusion model and other Generative AI frameworks. These frameworks are source and helped the authors make their Medical Text-to-Image Generation Using Generative AI system. The authors used image datasets that are available to everyone. These datasets helped the authors make and test their research. The authors are thankful for Medical Text-, to-Image Generation Using Generative AI and the help they got from the Stable Diffusion model and other Generative AI frameworks.

REFERENCES

1. R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Omer wrote a paper called "High-Resolution Image Synthesis with Latent Diffusion Models" for the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. It was published in 2022.
2. A. Radford, J. W. Kim, C. Hallacy and others published a paper called "Learning Transferable Visual Models From Natural Language Supervision" in the Proceedings of the International Conference on Machine Learning. This was in 2021.
3. A. Nichol and P. Dhariwal wrote a paper called "Improved Denoising Diffusion Probabilistic Models" for the Proceedings of the International Conference on Machine Learning. It was published in 2021.

4. P. Dhariwal and A. Nichol published a paper called "Diffusion Models Beat GANs on Image Synthesis" in Advances in Neural Information Processing Systems. This was in 2021.
5. A. Saharia, W. Chan, S. Saxena and others wrote a paper called "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding" for Advances in Neural Information Processing Systems. It was published in 2022.