AI Based Text to Image Generation

# Major Project Synopsis Report

*Submitted in partial fulfilment of the requirement of the degree of*

**BACHELOR OF TECHNOLOGY**

**in**

**CSE with Specialization AI & ML**

*to*

**K.R Mangalam University**

*by*

# Aditya Panwar (2201730074) Raghav Mathur (2201730091)

Under the supervision of

# Vishwanil Suman



Department of Computer Science and Engineering School of Engineering and Technology

K.R Mangalam University, Gurugram- 122001, India August 2025

# INDEX

|  |  |  |
| --- | --- | --- |
| 1. | Abstract | Page No. |
| 2. | Introduction (description of broad topic) |  |
| 3. | Motivation |  |
| 4. | Literature Review |  |
| 5. | Gap Analysis |  |
| 6. | Problem Statement |  |
| 7. | Objectives |  |
| 8. | Tools/platform Used |  |
| 9. | Methodology |  |
| 10. | References |  |

**ABSTRACT**

Artificial Intelligence (AI) has recently made remarkable progress in the field of generative models, particularly in text-to-image synthesis. Text-to-image generation enables the automatic creation of images from natural language descriptions by combining Natural Language Processing (NLP) with Computer Vision techniques. With the help of advanced models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models like stable diffusion. AI systems can now generate high-quality, realistic, and contextually accurate images.

This project explores the design and implementation of AI-based text-to-image generation systems, highlighting their architecture, performance, and applications. The study focuses on how these models learn semantic relationships between words and visual features, and how they can be fine-tuned for improved accuracy and efficiency. Applications of text-to-image models range from creative industries like art and entertainment to practical uses in education, advertising, product design, and healthcare visualization.

The aim of this work is to develop a prototype that not only generates visually appealing images from text but also ensures contextual relevance and computational efficiency. By addressing existing challenges such as bias, ethical concerns, and computational cost, the project seeks to contribute toward more reliable and responsible AI-driven image generation.

In diffusion models, stable diffusion is very advanced model. It progressively denoises random noise guided by text prompts to produce high-quality, photorealistic, and contextually accurate images. It operates in a latent space rather than pixel space, which makes it computationally efficient and suitable for running on consumer-level GPUs. Being open-source, Stable Diffusion has empowered developers, researchers, and creators to build customized applications for art, design, and content generation.

Another image generation model; DALL·E, developed by OpenAI, it leverages transformer-based architectures to create novel images from natural language descriptions. It extends the principles of GPT models by interpreting textual prompts and mapping them into coherent visual outputs. Unlike earlier generative models, DALL·E can combine unrelated concepts, generate imaginative scenes, and produce contextually aligned visuals that match user-provided text.

***KEYWORDS: Text-to-Image, Artificial Intelligence, Diffusion Models, GANs, Deep Learning***

# INTRODUCTION

Artificial Intelligence (AI) is revolutionizing the way humans interact with technology by enabling machines to understand, interpret, and generate human-like outputs. Among the many breakthroughs in this domain, **text-to-image generation** has emerged as one of the most fascinating advancements, as it effectively bridges the gap between natural language and computer vision. This technology allows users to generate realistic and contextually accurate images by simply providing a textual description. For example, a prompt such as *“a cat wearing a red hat sitting on a chair”* can lead an AI model to create a visually coherent image that aligns with the given description.

Recent breakthroughs, including **DALL·E, Stable Diffusion, and MidJourney**, have demonstrated the potential of integrating **natural language processing (NLP), computer vision, and generative modeling** into unified systems. These models leverage large-scale training datasets, attention mechanisms, and deep neural networks to synthesize high-resolution, creative, and context-aware visuals.

Their ability to capture fine-grained textual details and translate them into meaningful visual features establishes text-to-image generation as a critical research domain in AI and Machine Learning.

The growing applications of this technology span across a wide range of industries, highlighting its transformative potential in the digital era. In **digital art and design**, AI-based text-to-image systems act as creative assistants, enabling artists to rapidly visualize ideas and generate unique artwork without the need for manual illustration. In the **advertising and marketing industry**, these models are used to create customized visuals tailored to specific campaigns, reducing production time and cost. In **education**, text-to-image generation provides interactive visual aids, making complex topics easier to understand and more engaging for learners. Similarly, in **healthcare**, the technology can assist in generating medical illustrations and simulations from textual descriptions, supporting both professional training and patient education. The **entertainment and gaming industries** are also leveraging these tools to design characters, environments, and storyboards, accelerating the creative process.

# MOTIVATION

The motivation behind this project arises from the increasing demand for intelligent systems capable of transforming abstract ideas into tangible visuals without requiring advanced artistic or technical expertise. Traditionally, the creation of high-quality visuals has demanded considerable time, effort, and specialized skills in fields such as graphic design or digital illustration. With the advancement of text-to-image generation, this gap can now be bridged by enabling individuals to describe their vision in natural language and obtain corresponding images in real time. This not only enhances creativity but also democratizes access to design and visual storytelling.

For **content creators, advertisers, and educators**, such systems offer a cost-effective and time-efficient means of generating customized visuals tailored to specific needs. In industries such as **architecture, product design, and gaming**, text-to-image AI significantly accelerates the prototyping process, allowing designers to rapidly explore multiple design variations before finalization. This rapid iteration reduces development cycles, fosters innovation, and enhances productivity.

In addition, text-to-image technology has great potential in advancing **accessibility and inclusivity**. It can, for example, assist visually impaired individuals by converting text into meaningful visual representations, bridging the gap between textual information and perception. In educational contexts, it can simplify complex concepts through interactive and personalized visualizations, thereby improving engagement and learning outcomes.

The diverse applicability of this technology across multiple domains underscores the importance of continued research and development in text-to-image generation. By addressing challenges such as **semantic accuracy, ethical considerations, and computational efficiency**, this project seeks to advance the field and contribute toward making AI-driven creativity more accessible, reliable, and impactful.

# LITERATURE REVIEW

DALL·E (OpenAI, 2021):

* DALL·E, developed by OpenAI, demonstrated the power of transformer-based architectures in generating novel images from natural language descriptions. Extending the GPT framework, DALL·E was trained on a massive dataset of text–image pairs and showcased the ability to generate contextually coherent and imaginative visuals. It introduced the concept of compositionality in AI, where unrelated objects or abstract ideas could be combined into meaningful images. This model marked a significant milestone in bridging natural language understanding with visual creativity.

Stable Diffusion (Stability AI, 2022):

* Stable Diffusion introduced a new class of **diffusion models** that progressively denoise random noise into high-resolution and photorealistic images, guided by text prompts. Unlike earlier pixel-based models, Stable Diffusion operates in **latent space**, making it computationally efficient and accessible on consumer-grade GPUs. Its open-source release democratized text-to-image research and allowed developers to create customized applications in fields such as digital art, design, and education. It stands out for its scalability, flexibility, and balance between realism and creativity.

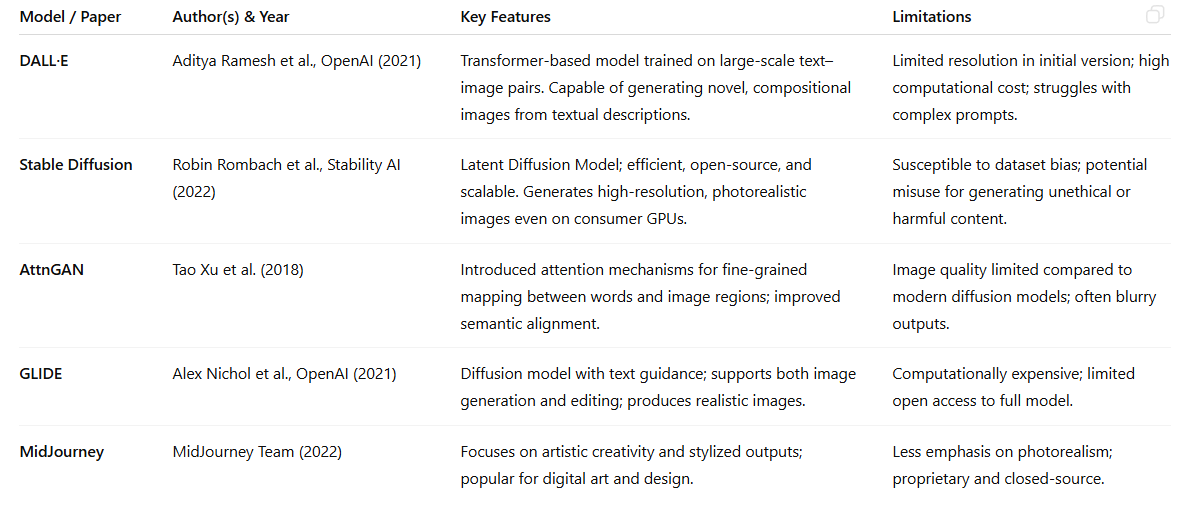
AttnGAN (Xu et al., 2018):

* Attentional Generative Adversarial Networks (AttnGAN) introduced the use of **attention mechanisms** to improve the alignment between textual attributes and visual features. By focusing on different words or phrases within the input text during image synthesis, AttnGAN achieved fine-grained control over the generation process. This advancement significantly improved the semantic accuracy of generated images compared to earlier GAN-based models, making it a foundational approach for subsequent research in text-to-image generation.

MidJourney (2022):

* MidJourney is an AI system that emphasizes **artistic creativity and aesthetics** rather than pure realism. Through a combination of proprietary generative models and iterative refinement, it enables the production of highly stylized, imaginative, and visually striking artwork from textual prompts. MidJourney gained popularity among artists and designers for its ability to create unique digital art, demonstrating the role of AI not only as a tool for realism but also as a collaborator in human creativity.

Table: Sample of Literature Review Table



# GAP ANALYSIS:

* **Lack of Contextual Accuracy:**  
   Many models struggle with understanding complex or abstract textual prompts. This often results in outputs that partially match the description but fail to capture fine-grained details or nuanced semantics.
* **Ethical Concerns:**  
   The same technology that enables creativity can also be misused for generating **deepfakes, misinformation, or inappropriate content**. This raises significant ethical and societal challenges regarding the safe and responsible use of such systems.
* **High Computational Requirements:**  
   Training and running advanced models typically demand substantial GPU resources, making them less accessible for individuals or organizations with limited computational power. This also restricts deployment in real-time or resource-constrained environments.
* **Limited Domain-Specific Customization:**  
   Most existing systems are trained on large, general-purpose datasets. While this allows for broad capabilities, they often lack precision in **specialized fields** such as medical imaging, industrial design, or scientific visualization, where domain-specific accuracy is crucial.

# PROBLEM STATEMENT

Although current text-to-image generation models are highly advanced and capable of producing realistic visuals, they still suffer from several critical limitations. Most models require **substantial computational resources**, restricting their accessibility to only well-equipped research labs or organizations with high-performance GPUs. Additionally, these systems often **lack interpretability**, making it difficult to understand how textual descriptions are translated into visual features. This poses challenges in debugging, improving, and trusting the outputs of such models.

Another challenge lies in their ability to handle **complex or abstract descriptions**. While models perform well on simple prompts, they frequently misinterpret fine-grained details or fail to capture contextual nuances, leading to outputs that do not fully align with user expectations. Moreover, existing systems are often designed for **general-purpose applications** and lack adaptability to **domain-specific contexts** such as medical visualization, industrial prototyping, or educational illustrations, where precision and accuracy are vital.

Therefore, there is a pressing need for the development of **efficient, reliable, and context-aware AI models** that not only generate accurate and semantically aligned images but also operate in a **resource-friendly and interpretable manner**. Such advancements would ensure broader accessibility, promote responsible use, and enable the application of text-to-image technology across diverse industries.

# OBJECTIVES

**Sample Objectives**

The primary objective of this project is to design and implement an **AI-based text-to-image generation system** that can effectively translate natural language descriptions into contextually accurate and visually coherent images. The specific objectives include:

* **To design and develop a prototype AI-based text-to-image generation system** that demonstrates the integration of natural language processing and computer vision techniques.
* **To explore state-of-the-art generative approaches**, including **Generative Adversarial Networks (GANs), Diffusion Models, and Transformer-based architectures**, inorder to identify the most effective methods for high-quality image synthesis.
* **To enhance contextual accuracy** by improving the semantic alignment between textual input and the generated output, ensuring that even complex or abstract prompts are represented effectively.
* **To evaluate system performance** using qualitative and quantitative measures, such as image resolution, realism, user satisfaction, and semantic similarity metrics.
* **To investigate real-world applications** of text-to-image AI in domains such as **education, entertainment, digital art, advertising, and product design**, thereby highlighting its potential impact across industries.

# Tools/Technologies Used

* For the development and implementation of this project, a combination of programming languages, deep learning frameworks, and pre-trained models will be utilized. The selected tools are widely adopted in the Artificial Intelligence and Machine Learning community, providing flexibility, scalability, and strong community support.
* **Programming Language: Python**  
   Python has been chosen as the primary programming language due to its simplicity, readability, and extensive ecosystem of libraries for Artificial Intelligence, Machine Learning, and Computer Vision. Its versatility allows seamless integration of model training, evaluation, and deployment.
* **Libraries / Frameworks: PyTorch, TensorFlow, Hugging Face Transformers**
  1. *PyTorch* and *TensorFlow* will be used for building and fine-tuning deep learning models, offering GPU acceleration and robust support for generative tasks.
  2. *Hugging Face Transformers* provides access to state-of-the-art pre-trained models and NLP tools, which will be integrated to improve text understanding and semantic alignment with generated images.
* **Pre-trained Models: Stable Diffusion, DALL·E-mini, AttnGAN**  
   These models will serve as baselines for experimentation and comparison.
* *Stable Diffusion* will be explored for high-resolution, resource-efficient image generation.
* *DALL·E-mini* provides a lightweight alternative for creative outputs.
* *AttnGAN* offers insight into attention-based alignment between words and image regions.
* **Supporting Tools: OpenCV, Streamlit/Flask, CUDA-enabled GPUs**
* *OpenCV* will be used for image preprocessing and handling.
* *Streamlit/Flask* will be employed to build an interactive user interface (UI) for generating and displaying images from text prompts.
* *CUDA-enabled GPUs* are essential for accelerating model training and inference, ensuring efficient performance.

# METHODOLOGY

The proposed project will follow a systematic methodology to design, develop, and evaluate an AI-based text-to-image generation system. The steps include:

* **Data Collection:**

Large-scale text–image datasets such as **COCO (Common Objects in Context)**, **Conceptual Captions**, and **LAION-5B** will be utilized. These datasets provide millions of paired text–image samples, enabling the model to learn semantic relationships between textual descriptions and their corresponding visuals.

* **Preprocessing:**
  1. **Text Processing:** Text descriptions will be cleaned, tokenized, and converted into embeddings using transformer-based language models (e.g., CLIP or BERT).
  2. **Image Processing:** Images will be normalized, resized, and augmented to ensure consistency and robustness during training.
* **Model Selection:**

Multiple architectures will be explored:

* 1. **Diffusion Models** (e.g., Stable Diffusion) for generating high-resolution, realistic images.
  2. **GAN-based Architectures** (e.g., AttnGAN) to enhance text–image alignment.
  3. **Transformer-based Models** for improved semantic understanding of complex prompts.
* **Training / Fine-Tuning:**
  1. Models will be trained on paired text–image datasets.
  2. Fine-tuning will be carried out to improve **semantic alignment** and reduce inconsistencies in generated outputs.
  3. Transfer learning will be applied where applicable to optimize training efficiency.
* **Evaluation:**

The system will be evaluated using both **quantitative metrics** and **qualitative analysis.**

* 1. **Fréchet Inception Distance (FID):** To measure realism of generated images.
  2. **CLIP Similarity Score:** To evaluate semantic alignment between text and image.
  3. **User Feedback:** To assess usability, creativity, and overall quality of results.
* **Deployment:**

An interactive **prototype application** will be developed using **Streamlit or Flask**, allowing users to input textual prompts and view generated images in real time. This deployment will demonstrate practical usability and highlight real-world applications of the system in areas such as education, entertainment, and design.

**REFERENCES:**

1. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., … & Sutskever, I. (2021). *Zero-shot text-to-image generation*. Proceedings of the 38th International Conference on Machine Learning (ICML). OpenAI.
2. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., … & Chen, M. (2021). *GLIDE: Towards photorealistic image generation and editing with text guidance*. arXiv preprint arXiv:2112.10741.
3. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-resolution image synthesis with latent diffusion models*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10684–10695.