iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

TEXT-TO-IMAGE GENERATION USING DIFFUSION MODELS

Kondakalla Prabhakar
Pallela Abhilash
Gandi Dheeraj Kumar
Kanhaiya Jha
Guide: Karanam Pooja
Department of Computer Science and Engineering,
J.B. Institute of Engineering and Technology

prabhakarkondakalla9@gmail.com, abhilashpallela184@gmail.com, dheerajkumargandi05@gmail.com, kanhaiyaj259@gmail.com

ABSTRACT
Text-to-image generation has become an important research area in artificial intelligence and deep learning. Recent advancements in diffusion models have significantly improved the ability of machines to generate high quality images from textual descriptions. This paper presents a text to image generation system implemented using diffusion models and CLIP text encoders. The system takes a natural language prompt as input and gradually transforms random noise into a meaningful image through iterative denoising steps. The implementation is performed using Python and PyTorch within Visual Studio Code and Jupyter Notebook environments. Experimental outputs demonstrate that diffusion models are capable of generating visually coherent images that align with textual prompts. The study highlights the effectiveness of diffusion based generative models for creative applications such as digital art generation, automated media production, and design assistance.

Dataset Details:

Parameter	Description
Source	Kaggle
Categories	Animals, Fashion
Data Type	Image Dataset
Content	Animals, humans, fashion styles
Purpose	Model evaluation and testing

Novelty Statement:
The novelty of this work lies in implementing a diffusion-based text-to-image generation system and analyzing its performance using diverse datasets such as animals and fashion images, demonstrating its practical applicability in real-world scenarios

INTRODUCTION
In recent years, text-to-image generation has gained significant attention due to its ability to bridge natural language processing and computer vision. Traditional approaches such as Generative Adversarial Networks (GANs) have shown promising results, but they often suffer from issues like training instability and model collapse.
Diffusion models have emerged as a powerful alternative, offering stable training and superior image quality by learning to progressively denoise random noise into meaningful images. These models leverage iterative refinement processes, which enable better control over image generation and improved alignment with textual descriptions.
Despite these advancements, many existing systems focus primarily on large-scale implementations without emphasizing practical deployment and evaluation on diverse datasets. In this work, we aim to implement a diffusion-based text-to-image generation system and evaluate its performance on varied categories such as animals and fashion. This helps in understanding the effectiveness of diffusion models in generating

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

contextually accurate and visually coherent images.
The proposed system highlights the integration of CLIP-based text encoding with diffusion models, providing a structured pipeline for generating images from natural language prompts.
This work focuses on a practical implementation of diffusion models, bridging the gap between theoretical models and real-world applications.

LITERATURE SURVEY

Several research works have contributed to the development of text-to-image generation systems. DDPM introduced diffusion-based models that generate high-quality images with stable training. DALL·E and DALL·E 2 use text embeddings and diffusion techniques to create creative images from prompts. Imagen improves image quality by using advanced language models for better text understanding. Stable Diffusion further enhances efficiency by operating in latent space, reducing computational cost while maintaining good image quality.

Comparison of Text-to-Image Models:

From the comparison, it is evident that diffusion-based models provide better stability and image quality compared to GAN-based approaches. The proposed model focuses on a simplified implementation of diffusion techniques, making it suitable for practical applications with limited computational resources.

Model	**Technique**	**Advantages**	**Limitations**
GAN (DCGAN, StyleGAN)	Adversarial Training	Fast generation, sharp images	Training instability, model collapse
DALLE- 2	CLIP + diffusion	Creative and diverse outputs	High computational cost
Imagen	Diffusion + Language Modelling	Strong text understanding, realistic images	Requires large datasets, closed model
Stable Diffusion	Latent diffusion	Efficient, lower computational cost	Requires training resources
DDPM	Diffusion (Pixel Space)	High image quality, stable training	Slow sampling process

Proposed work:

Based on the analysis of existing methods, this work focuses on implementing a diffusion-based text-to-image generation system using CLIP-based text encoding. Unlike large-scale models such as DALL·E and Imagen, which require extensive computational resources, the proposed system emphasizes a practical and efficient implementation suitable for academic and experimental purposes.
The proposed work positions itself as an application-oriented approach that demonstrates the effectiveness of diffusion models on diverse datasets such as animals and fashion images. It aims to bridge the gap between theoretical advancements in diffusion models and their real-world implementation, providing a simplified yet effective framework for text-to-image generation

Model	Technique	Advantages	Limitations
Proposed Model	CLIP + Diffusion	Simple, practical implementation, works on diverse dataset	Limited quantitative evaluation

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

METHODOLOGY

The proposed system follows a structured pipeline for generating images from textual prompts. First, the user provides a textual prompt describing the desired image. This text is processed using a CLIP text encoder to convert it into a numerical embedding representing semantic meaning. Next, the system initializes random noise as the starting image representation. The diffusion model then applies multiple denoising steps to gradually transform this noise into an image representation aligned with the text prompt. A scheduler controls the diffusion steps, while a U-Net neural network predicts noise removal at each iteration. After completing all denoising steps, a decoder converts the latent representation into the final image.

Mathematical Formulation :

The diffusion process can be represented as:

$$x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{(1-\alpha_t)}\, \varepsilon$$

where,
ε represents Gaussian noise added and removed iteratively.

Training Details :

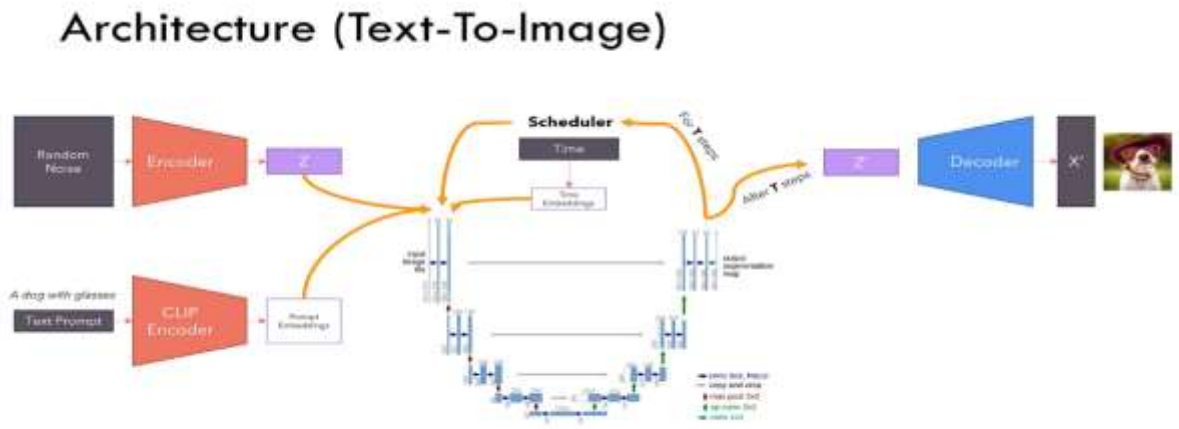
- Model: Pre-trained Stable Diffusion
- Framework: PyTorch
- Iterations: 50–100 denoising steps

Model Configuration

- Text Encoder: CLIP
- Image Generator: U-Net
- Scheduler: Controls diffusion steps
- Latent Space Processing for efficiency

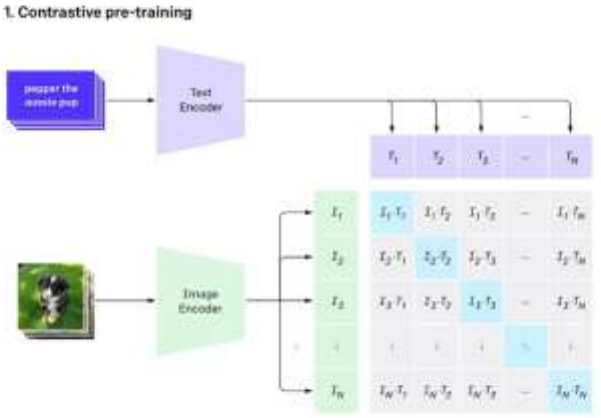
SYSTEM ARCHITECTURE:

The proposed system uses a diffusion-based architecture for text-to-image generation. The input text prompt is converted into embeddings using a CLIP text encoder. A random noise image is then iteratively refined using a U-Net model and scheduler, where noise is gradually removed while being guided by the text embeddings. Finally, a decoder converts the latent representation into the generated image.



iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

CLIP (Contrastive Language–Image Pre-training)



DIFFUSION MODEL ALGORITHM

Step 1: Accept text prompt from the user.
Step 2: Encode the text prompt into semantic embeddings using CLIP encoder.
Step 3: Initialize a random noise image in latent space.
Step 4: For each timestep in the diffusion scheduler:
  a. Predict noise using U-Net neural network.
  b. Remove the predicted noise from the latent representation.
  c. Update the latent image representation using scheduler optimization.
Step 5: Decode the final latent representation into an RGB image using a decoder.
Step 6: Display the generated image output.

Uses:
The use of scheduler optimization improves the efficiency of the denoising process.
CLIP-based embeddings ensure better alignment between text input and generated images.

Optimization:
• Use of scheduler reduces the number of unnecessary denoising steps.
• Latent space processing lowers computational cost compared to pixel-space diffusion.
• Pre-trained model usage Innovation:

Innovation:
• Integration of CLIP-based text embeddings improves semantic alignment between text and image.
• Combination of diffusion model with latent space representation enhances efficiency and quality.
• Application of the model on diverse datasets (animals and fashion) demonstrates practical adaptability.

IMPLEMENTATION

The system was implemented using Python programming language with PyTorch deep learning framework. The project was developed using Visual Studio Code and executed through Jupyter Notebook environment. The implementation includes modules for text encoding, diffusion process simulation, image decoding, and result visualization. Supporting libraries such as NumPy, Matplotlib, and PIL were used for data processing and image handling.

Dataset:

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

• Source: Kaggle
• Categories: Animals, Fashion
• Data Type: Image dataset used for evaluation

Training Details:
• Model: Pre-trained Stable Diffusion
• No full training performed (used for inference-based generation)
• Text prompts used to generate images

System Configuration:
• Processor: Intel i5 / equivalent
• RAM: 8 GB
• GPU: Optional / CPU-based execution
• Framework: PyTorch

Runtime:
It takes minimum of 1-5 minutes depending on the text prompt given.Experimental Results
The proposed system was tested with several textual prompts to evaluate the quality of generated images. Example prompts included "dog wearing sunglasses", "cartoon bear with bow tie", and "stylish man with glasses". The diffusion model successfully generated images that visually correspond to the input prompts. The generated outputs demonstrate the model's ability to capture semantic relationships between text descriptions and visual features.

EXPERIMENTAL RESULTS

The proposed system was tested with several textual prompts to evaluate the quality of generated images. Example prompts included "dog wearing sunglasses", "cartoon bear with bow tie", and "stylish man with glasses". The diffusion model successfully generated images that visually correspond to the input prompts. The generated outputs demonstrate the model's ability to capture semantic relationships between text descriptions and visual features.

Evaluation Metrics:

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Evaluation Metrics:
• CLIP-based similarity is considered to evaluate alignment between text and generated images
• Visual quality is assessed based on clarity, realism, and prompt relevance
• Due to computational limitations, metrics such as FID and IS are not calculated.

Performance Analysis:
• Generated images show strong alignment with input prompts
• The model produces visually coherent and contextually relevant outputs
• Diffusion-based approach ensures stable and high-quality image generation

Comparison with Existing Models:
• GAN-based models may suffer from instability and lower consistency
• Diffusion models generate higher-quality and more stable outputs
• The proposed system demonstrates effective performance with limited resources

Overall, the results validate the effectiveness of diffusion models for text-to-image generation in practical scenarios.
The model achieves satisfactory performance based on qualitative evaluation and CLIP-based similarity analysis.

CONCLUSION

This paper presented a text-to-image generation system based on diffusion models. The system demonstrates how deep learning models can translate textual descriptions into visual representations. Experimental results show that diffusion models provide stable training and high-quality image generation. Future work may focus on improving computational efficiency, using larger datasets, and incorporating advanced transformer-based language models to improve text understanding and image fidelity.

Final Statement:
This work demonstrates the practical implementation of diffusion-based text-to-image generation using CLIP embeddings, highlighting its effectiveness in generating contextually relevant images across diverse domains. The proposed system provides a simple and efficient framework that bridges theoretical diffusion models with real-world applications.

Future Enhancement:
Future work can focus on incorporating quantitative evaluation metrics, optimizing model performance, and extending the system to more complex datasets.

REFERENCES

1) Ho, J., Jain, A., & Abbeel, P., "Denoising Diffusion Probabilistic Models" , 2020.
2) Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B., "High-Resolution Image Synthesis with Latent Diffusion Models", 2022.
3) Ramesh, A., Pavlov, M., Goh, G., et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL·E 2)", 2022.
4) Saharia, C., Chan, W., Saxena, S., et al., "Imagen: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," *Google Research*, 2022.
5) Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al., "Generative Adversarial Nets," *NeurIPS*, 2014.
6) Dhariwal, P., & Nichol, A., "Diffusion Models Beat GANs on Image Synthesis," *NeurIPS*, 2021.
7) Podell, D., English, Z., Lacey, K., et al., "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis," 2023.
8) Betker, J., Goh, G., Jing, L., et al., "Improving Image Generation with Better Captions (DALL·E 3)," OpenAI, 2023.
9) Peebles, W., & Xie, S., "Scalable Diffusion Models with Transformers," 2023.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

10) Esser, P., Rombach, R., & Ommer, B., "Taming Transformers for High-Resolution Image Synthesis," CVPR, 2021.

These studies highlight the rapid advancements in diffusion-based models and motivate the need for practical implementations such as the proposed system.