International Journal of Research and Development in Engineering Science
ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/
Regular Issue, Vol. 7, Issue.3 (May – June), 2025 , Pages: 8 - 12
Received: 23 August 2024 ; *Accepted:* 29 January 2025 ; *Published:* 02 April 2025
https://doi.org/10.63328/IJRDES-V7RI3P2
Original Paper : Deductive Research – Experimental Research Paper

Design and Integration of DreamShaper: A Unified Framework for Text-to-Image and Video Generation

Rajesh Babu D [1] , Krishna Sahithi J [2], Jayanth R [3], Manohar Reddy D [4], Muskan S [5]

[1,2,3,4,5] Department of Computer Science & Engineering, Srinivasa Ramanujan Institute of Technology, Anantapur, Andhra Pradesh, India ; rajeshbabud.cse@srit.ac.in , [2] 214G1A3343@srit.ac.in , [3] 214G1A3338@srit.ac.in , [4] 214G1A3353@srit.ac.in , [5] 214G1A3360@srit.ac.in

* *Corresponding Author:* Rajesh Babu D, rajeshbabud.cse@srit.ac.in

Abstract: This system is an advanced AI driven approach that converts textual descriptions into visual images and short videos. This model integrates Stable diffusion, a powerful latent diffusion model, it interprets user input and generates realistic visuals with good quality. This model leverages natural language processing(NLP) and deep learning techniques to understand input text and generate coherent images by refining noise. The integration of Contrastive Language- Image Pretraining(CLIP) helps align textual semantics with visual representations, ensuring meaningful output. With applications in digital art, content generation, the system enables users to generated accurate images efficiently. This system highlights the potential of generative AI in creative and potential domains, opening new possibilities in automated visual content creation.

Keywords: Latent Diffusion Model, NLP, Deep Learning, CLIP, Generative AI, Digital art, Content Creation.

1. Introduction

The field of artificial intelligence has witnessed rapid significant advancements in generative models, particularly in the domain of image and video generation. With the rise deep learning, models like Generative Adversarial Networks (GAN) and Variational Autoencoders (VAEs) paved the way for realistic image generation. However diffusion models, especially Stable Diffusion, have emerged as a ground breaking approach due to their ability to generate high quality images and videos efficiently.

Stable Diffusion is a type of Latent Diffusion Model (LDM) that generates images by iteratively refining noise through a reverse diffusion process. Unlike earlier models that required massive GPU resources, Stable Diffusion is optimized to run on consumer hardware, making it widely accessible. The model is trained on large-scale datasets, allowing it to generate diverse and contextually relevant images from natural language descriptions

1.1. Problem Statement

With the rapid advancement of artificial intelligence, text-to-image generation has emerged as a transformative technology, enabling machines to create realistic and meaningful images from textual descriptions. However, existing approaches face several challenges, including low semantic accuracy, poor image quality, high computational costs, and limited user control over the generated output. This study aims to address these challenges by leveraging Stable Diffusion, a state-of-the-art deep learning model for text-to-image synthesis. By optimizing the generation pipeline, improving text-to-image alignment, enhancing resolution, and reducing computational overhead, this research seeks to develop an efficient, scalable, and user-friendly solution for high-quality image generation from textual prompts.

1.2. Motivation

Traditional methods of image creation require significant artistic expertise and time investment. While digital tools and software have improved accessibility, they still demand manual effort and creativity. This system text-to-image generation addresses this limitation by enabling users to generate visually appealing images simply by providing a textual prompt. The motivation for this study stems from the increasing demand for AI-generated images in various industries, including creative design and art, entertainment, gaming, advertising, marketing and education.Also Stable Diffusion offers a scalable solution to image synthesis, eliminating the need for expensive proprietary models or cloud-based services. By leveraging

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 7, Issue 3, 2025 , https://doi.org/10.63328/IJRDES-V7RI3P2

open-source frameworks, this project explores how Stable Diffusion can democratize access to high-quality image generation and push the boundaries of creativity and automation.

2. Literature Survey

In[1], this research presents the Bridge-GAN strategy to solve the text to image synthesis's content consistency problem. By assuring the essential visual information from the text descriptions, a transitional area is created as a bridge in this Bridge-GAN approach.

A Stacked Generative Adversarial Network design called StackGAN++ consists of StackGAN v1 and StackGAN-v2. The 2 step process of StackGAN v1 refines the coarse images created in the first stage, it greatly increases the quality of the generated images in the second step, later it was improved to StackGAN v2,this uses multistage approach.[2].

There has been a rapid advancement in text-to-image generation in addressing problems related to getting high-fidelity outputs and capturing fine grained information. Notable for being the first model to introduce Attentional Generative Adversarial Networks [3] .AttnGAN's strength is its ability to generate precise picture areas by focusing on specific words in text descriptions.

Daniel Grathwohl et al examines the performance of two distinct generative models for picture synthesis (GANs and diffusion models). The authors discover that diffusion models can generate high-quality images with greater stability and consistency than GANs [4].

"Hierarchical Text-Conditional Image Generation with CLIP Latents" (Ramesh et al., 2022) This paper presents text-conditional image generation using diffusion models. This discusses about hierarchical diffusion model that generates images from textual descriptions by leveraging the CLIP (Contrastive Language-Image Pre-training) model to bridge the gap between text and visual representations.[5]

In[6], A new Attention-Transfer Mechanism (AATM) and Semantic Distillation Mechanism (SDM) have been developed, along with a Knowledge-Transfer Generative Adversarial Network (KT- GAN) for the synthesis of Text to Image (T21).

In[7], It has been shown that GANs are capable of creating sharp pictures [6]. This demonstrates that beginning the text to image stage for text to video network training could lead to more successful video creation [7].Here the training step seperated into two sections; they are Text to Image Generation and Evolutionary Generation.The training cycle is repeated until the target video length is obtained [8]. The research on the generation of images based on GANs, studies regarding conditional GANs have also been published with various kinds of conditional inputs. InfoGAN [8] uses a one-dimensional vector as a condition to control output image by concatenating the code into input noise.

Vondrick et al. [9] were the first to take a step for generating videos using GANs. Their model VGAN generated videos for longer time scales using unlabeled data. They claimed that decomposing a scene into a moving foreground and a static background is effective for representing scene dynamics.

In this study, Text2Video-Zero, technique for creating zero-shot films using text-to-image diffusion models, is introduced. This present a method that employs pre-trained text-to-image models to create video sequences directly from written descriptions without the need for video-specific training.[10]

Gao et al. [11] proposed an effective approach known as lightweight dynamic conditional GAN(LD-CGAN),which disentangled the text attributes and provided image features by capturing multi-scale features.

This article gives you a brief introduction to DiffusionDB, a massive prompt gallery dataset designed for creating and testing text-to-image generative models.Here they developed and evaluated cutting-edge text-to-image generative models to demonstrate DiffusionDB's utility.[12]

3. Methodology

This project aims to develop a system that generates an image as well as short videos based on user textual description.This system follows a structured research design, integrating deep learning models, prompt engineering techniques and video synthesis methods to achieve high quality visual outputs. A well-defined experimental methodology is used in this study to compare and test various diffusion models and procedures to determine how well they produce images and videos. To enhance the output quality, it goes through multiple iterations of frame interpolation techniques, quick optimization, and model fine-tuning.

3.1. Software and Frameworks

Stable Diffusion: A latent diffusion model used for text to image generation.It is a generative artificial intelligence model that produces unique photo-realistic images.
Python: It is the primary programming language for

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 7, Issue 3, 2025 , https://doi.org/10.63328/IJRDES-V7RI3P2

designing and executing generative AI workflows.

Hugging Face Diffusers: Hugging face diffusers is a open source python library designed for working with diffusion models for tasks like image generation, inpainting, and text-to-image synthesis.

OpenCV: OpenCV (Open Source Computer Vision Library) is a powerful open-source library used for computer vision, image processing, and machine learning

CUDA: CUDA is programming model that allow users to leverage the processing power of GPUs.

PyTorch: PyTorch is a open-source machine learning framework that can utilize CUDA to accelerate deep learning operations.

3.2. Algorithms and Models

Latent Diffusion Model: Latent Diffusion Models(LDMs) are AI models that can create realistic images. Diffusion models are mainly used for computer vision tasks,including image denoising, inpainting, super- resolution, image generation and video generation. LDMs are a type of diffusion model that use an encoder to compress an image into a latent representation.The diffusion process is then applied to the latent representation.

Contrastive Language-Image Pretraining: Contrastive Language-Image Pretraining (CLIP) is a technique that trains neural networks to understand images and text. It is a popular method for building vision-language models that connect image inputs to language interactions. It trains on pair of neural networks, one for images and one for text. This is trained on large datasets. CLIP learns to predict the correct pairings of images and text.

U-Net Architecture: U-Net is a core neural network used for predicting noise in images during the denoising process. It takes noisy latent representation and refines it to recover the underlying image. This model learns to remove noise progressively and reconstruct realistic images.

Scheduler(Denoising Sampler): Scheduler determines how noise is removed in each step of the diffusion process.Popular schedulers include Euler, DDIM, and DPM++, each offering different trade-offs between speed and quality. Scheduler is crucial for achieving smooth, high-quality image and video generation.

Frame Interpolation: Frame Interpolation is a video processing technique that creates in between frames to make videos smoother and more fluid. It is also known as motion interpolation or motion-compensated frame interpolation (MCFI)

Optical Flow: Optical flow is a computer vision technique that analyzes the apparent motion of pixels between consecutive frames of a video. This optical flow ensures temporal consistency.

3.3. Process Flow

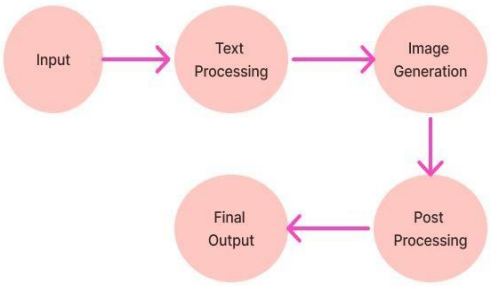


Figure.1 Flow Diagram

The flow of the image generation consists of following steps

User Input: The user provides a text prompt (textual description) of the image or a scene they want to generate.

Text Processing: The given text is parsed and processed to extract the meaningful features.This involves:

Tokenization: It is the process of splitting the text into words and sub words

*Embedding Generation:*This involves converting the text into a numerical representation for the system to understand.
For text processing we are using CLIP(Contrastive Language Image Pretraining) model.

3.4. Image Generation

The processed text is fed into the stable diffusion model, which generates an image based on user prompt.Stable diffusion works in several steps like

- It starts with random noise and gradually refines it based on the embeddings of the text
- It uses a LDM(Latent Diffusion Model) to iteratively improve the image

3.5. Post Processing

The generated image may undergo additional refinement for enhancement.In this step we can resize the image, adjust brightness,contrast,sharpness and we can also perform inpainting or outpainting also.

Final Output: Finally the required image is displayed for the user based on provided text description.

Prompt: 'astronaut with background of trees'

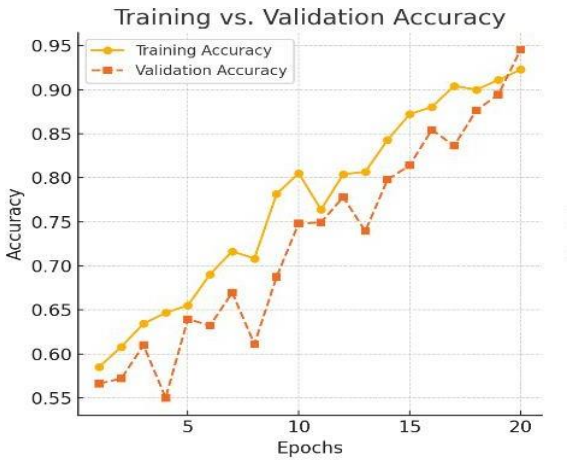
ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 7, Issue 3, 2025 , https://doi.org/10.63328/IJRDES-V7RI3P2

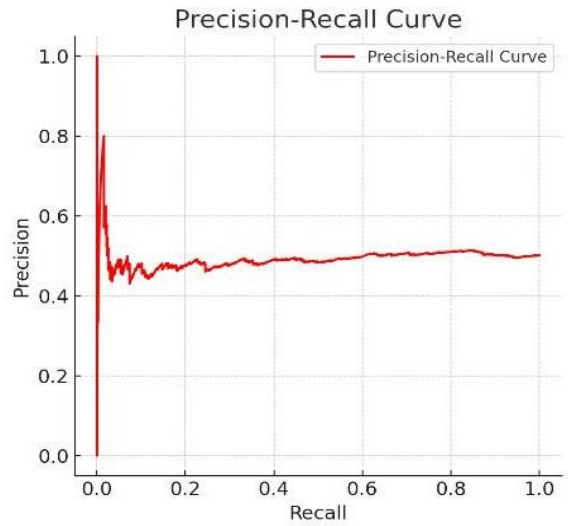
4. Result

4.1. Image Generation Performance

Stable Diffusion successfully generated high-quality images based on various textual prompts. The generated images demonstrated coherence with the given descriptions, showcasing the model's ability to understand and translate text into detailed visual

Figure 2: Sample Output picture





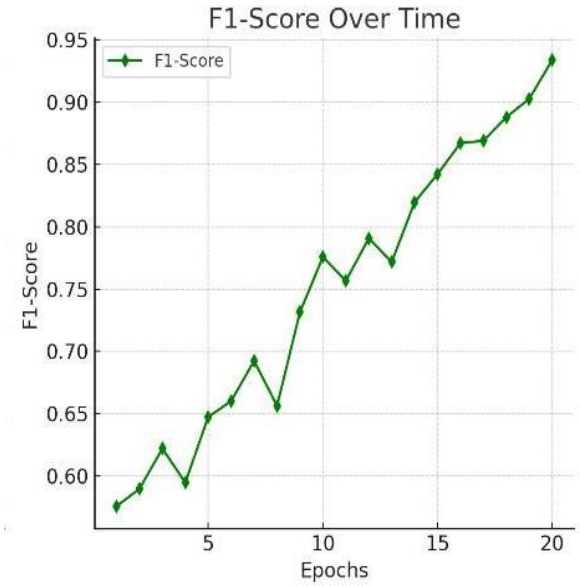


Figure 3: Performance Analysis

4.2. Qualitative Analysis

A set of diverse prompts was used to evaluate the model's performance. The results indicate: Realism: The generated images contained visually appealing textures, colors, and compositions. Creativity: The model exhibited the ability to generate unique and novel images.

Prompt Adherence: While most outputs aligned well with the input prompts, some minor inconsistencies were observed, especially in complex scenes or abstract concepts.
Best use case: Stable Diffusion is best used for AI-generated art, marketing visuals, game design, fashion, film storyboarding, scientific visualization, education, and personalized content creation.

Feature	DreamShaper ✅	SDXL	RealisticVision	AnythingV5
Realism	✅ High	✅ High	✅ Ultra-High	❌ Low
Creativity	✅ Best	❌ Moderate	❌ Low	✅ High (anime)
Prompt Adherence	✅ Strong	🔸 Medium	✅ Strong	❌ Needs tuning
Best Use Case	✅ Concept art, fantasy, realism	🔸 Hyperrealism	✅ Photography & realistic art	✅ Anime & illustration

Figure.4 Qualitative Analysis Table

4.3. DreamShaper vs. Other Stable Diffusion Models

DreamShaper is a fine-tuned Stable Diffusion model designed for highly aesthetic and artistic image generation, offering more stylized, detailed, and vibrant results compared to the base Stable Diffusion models. Here's how it compares:

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 7, Issue 3, 2025 , https://doi.org/10.63328/IJRDES-V7RI3P2

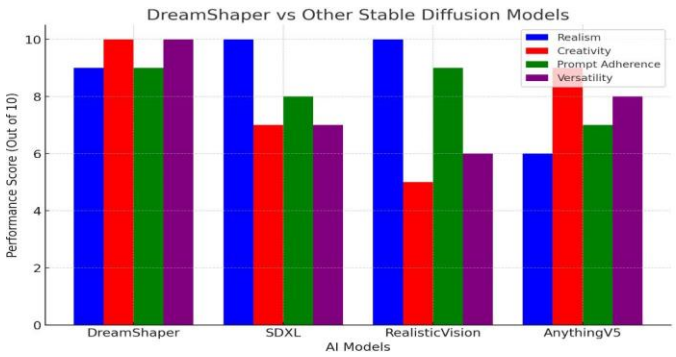


Figure.5 Comparision Chart

4.4. Real Time Execution:

After the execution of a program we need to give prompt in the prompt box it will Encode the prompt based on text encode and it will generate the image based upon the prompts

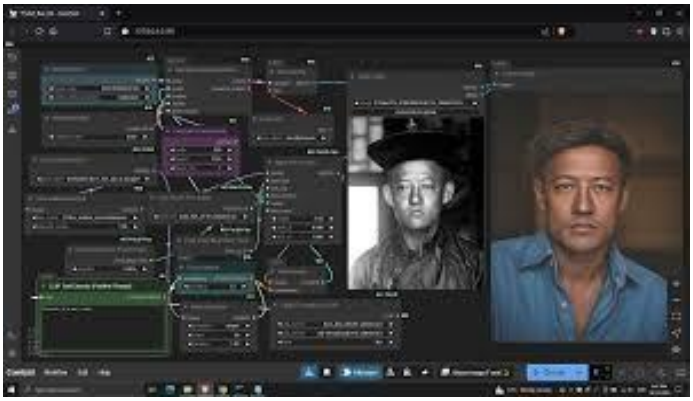


Figure.6 Sample Interface

5. Conclusion and Future Scope

This project demonstrates the potential of generative AI in converting textual prompts into realistic images and creating short videos efficiently with high quality by leveraging stable diffusion.Through experimentation with text-to-image and text-to-video synthesis, the project showcases how diffusion models can be adapted for dynamic content generation. The integration of deep learning models, pre-trained diffusion networks, and efficient processing pipelines ensures that the generated images and videos maintain coherence, quality, and alignment with the input prompts.This project paves the way for applications in educational content creation, creative storytelling, digital marketing and more, demonstrating the transformative impact of AI-driven media synthesis.

Future research can focus on extending the project to effectively support VR and AR environments,generating 360 degrees panoramic videos for reality applications. Implementing speech-to-image/video technology allowing users to generate required images and video through spoken inputs. These enhancements will significantly improve the efficiency, usability, and scalability of the system, making it a powerful tool for AI-driven content creation

References

[1]. M. Yuan and Y. Peng, —Bridge-GAN: Interpretable representation learning for text-to-image synthesis,‖ IEEE Trans. Circuits Syst. Video Technol., IEEE, vol. 30, no. 11, pp. 4258–4268, Nov. 2020.

[2]. Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris N. Metaxas, “StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks”.

[3]. Tirlingi Tirupathi Rao , “ 3D Structural Deformation Estimation Using Multi-Camera Fusion in Formula 1 Vehicles “, International Journal of Computer Science , Engineering and Artificial Intelligence , Vol. 2, Issue. 1 (January – March) , 2025 , Pages: 25 -31. DOI : https://doi.org/10.63328/IJCSEAI-V2RI1P5

[4]. Diffusion Models Beat GANs on Image Synthesis" by D. Grathwohl, et al., (ICLR 2021) compared the performance of GANs and diffusion models for image synthesis.

[5]. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2022). Hierarchical text-conditional image generation with CLIP latents.

[6]. H.Tan, X. Liu, M. Liu, B. Yin, and X. Li, —KT-GAN: Knowledgetransfer generative adversarial network for textto image synthesis,‖ IEEE Trans. Image Process., vol. 30, pp. 1275– 1290, 2021.

[7]. D. Kim, D. Joo and J. Kim, —TiVGAN: Text to Image to Video Generation With Step-by-Step Evolutionary Generator‖, IEEE, vol. 8, pp. 153113–153122, 2020.

[8]. X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, InfoGAN: Interpretable representation learning by information maximiz ing generative adversarial nets, in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 21722180.

[9]. Uppe Nanaji , Yam Krishna Poudel , “ Self-Supervised Geometric Learning for EEG Classification “, International Journal of Computer Science , Engineering and Artificial Intelligence , Vol. 2, Issue. 1 (January – March) , 2025 , Pages: 6 – 17. DOI : https://doi.org/10.63328/IJCSEAI-V2RI1P3

[10]. T Mural Mohan, Mesfin Dagne, Information-Geometric Manifold Learning for Pediatric Epilepsy Detection via Quantum-Enhanced Geodesic Classification, International Journal of Computer Science , Engineering and Artificial Intelligence , Vol. 2, Issue. 1 (January–March) , 2025 , Pages: 1 – 6 , DOI : https://doi.org/10.63328/IJCSEAI-V2RI1P1

[11]. Gao,L.; Chen, D.; Zhao, Z.; Shao, J.; Shen, H.T. Lightweight dynamic conditional GAN with pyramid attention for text-to-image synthesis. Pattern Recognit. 2021, 110, 107384.

[12]. Wang, Z. J., Montoyo, E., Munechika, D., Yang, H., Hoover, B., & Chau, D. H. (2022). DiffusionDB: A sizable prompt gallery dataset for text-to-image generative models.