Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 6s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

An Efficiency-Optimized Framework for Sequential Image Generation with Stable Diffusion Models

Dr. C. Beulah Christalin Latha[1], Dr. S.V. Evangelin Sonia[2], G. Linda Rose[3], Ben M. Jebin[4], Christhya Joseph[5], Dr. G. Naveen Sundar[6]

Karunya Institute of Technology and Sciences, Coimbatore.

Email ID: beulahchristudas@karunya.edu

Email ID: evangelinsonia.vs@gmail.com

* Email ID: lindarose@karunya.edu

Email ID: benmjebin@karunya.edu

Email ID: christhyajoseph@karunya.edu.in

Email ID: naveensundar@karunya.edu

Cite this paper as: Dr. C. Beulah Christalin Latha, Dr. S.V. Evangelin Sonia, G. Linda Rose, Ben M. Jebin, Christhya Joseph, Dr. G. Naveen Sundar, (2025) An Efficiency-Optimized Framework for Sequential Image Generation with Stable Diffusion Models. *Journal of Neonatal Surgery*, 14 (6s), 497-504

ABSTRACT

Diffusion models are a type of Generative Artificial Intelligence models that create data by adding noise to the data and gradually removing the noise to generate synthetic data. This research work focuses on visual content generation based on text input. We are using a computationally efficient variant of diffusion models namely, a stable diffusion model to create images from text. The novelty of this research work focuses on sequential image generation from text prompts. This approach has been successfully implemented to create visual storyboards, and diverse, contextually coherent images from text prompts. Hyperparameters have been fine-tuned to generate high-quality and context-aware images from text prompts.

Keywords: *Diffusion Models, Image Generation, Stable Diffusion, Text-to-Image, Storyboarding*

1. INTRODUCTION

There has been a significant surge in the use of Generative machine learning models in the recent past in various domains. Generative machine learning model is an advancing technology in the field of artificial intelligence. These models create synthetic data that resemble real data. Most machine learning models especially, deep learning models, need huge datasets for training themselves. Lack of real time data in critical domains such as healthcare, finance, and cybersecurity lead to challenges such as overfitting and reduced model accuracy. Researchers often find lack of data as a bottleneck when training machine learning models. This research addresses the issue of insufficient data in machine learning by generating synthetic data using generative models.

Diffusion models are a type of generative models that are efficient in generating synthetic data. They add noise to the input data and gradually remove the noise to create synthetic data that resembles real data. Traditional diffusion models are effective in generating high quality data, but they require multiple iterative steps for gradual denoising of data. This makes such models computationally expensive and slower than other generative models. This research uses stable diffusion models that are computationally more effective than traditional diffusion models for generating images from text.

Creating images from text prompts can be applicable to many domains. It enhances visualization of ideas in various fields and enables quick representation of concepts through illustrations, animations and storyboards. It enables visually challenged people to feel the aesthetic sense in art and experience visual storytelling through tactile graphics. Educationists use them to create visual representation for complex concepts to make understanding simpler. Visual models also help medical practitioners in better understanding of medical data, thus enabling a quicker and more accurate diagnosis. They are also applicable in various fields such as e-commerce, social media, and gaming. Text-to-image generations are more inclusive and they are also effective in bridging language and communication gaps.

2. LITERATURE SURVEY

A review on state-of-the-art literature reveals that generative models such as Generative Adversarial networks, Vector Quantized Variational Autoencoders, Diffusion based methods and Transformer models are commonly used for generating images from text. There are research gaps that exist in each of these models.

GAN-based Systems

Earlier text-to-image systems used GANs for generating images from text. GAN models such as StackGAN, AttnGAN, and Text2Image GAN are some of the commonly used GAN models in text-to-image applications. These models proved to be good in creating high-resolution images with good accuracy [1]. Some of the drawbacks identified in these models are:

Mode Collapse: This is a problem with GANs in which the generator produces a limited diversity in images leading to repetitive images in the output. They also do not cover the full text descriptions in many cases [2].

Training Instability: The training phase of GANs are computationally expensive and unstable. Tuning of hyperparameters is a tedious process especially in GANs. Small variations in hyperparameters may lead to extremely poor image quality or convergence failure [3].

Limited Coherence: GANs struggle in interpreting complex text descriptions with intricate scenes and maintaining a strong semantic alignment in such cases is challenging [4].

Resource Intensive: Training GANs is computationally expensive. They require processors with high performance and training also takes a lot of time. The training phase is highly time consuming in most of the cases [5].

VQ-VAE (Vector Quantized Variational Autoencoder) Based Systems

Text-to-image models such as DALL-E are VQ-VAE models. They generate high-quality images by learning from discrete latent representations of data. These models use Autoencoders for compression and decompression data. The input is passed through a narrow pipeline which enables removal of noise and regenerating the data results in noise-free output. However, the following drawbacks were observed in autoencoder-based models [6].

Lack of Fine-Grained Control: VQ-VAE models often fail to capture intricate details in text prompts and therefore, the output images do not exactly match the prompt strings.

Slow Inference: VQ-VAE architectures are slower in generating images especially when the prompts require to generate large images or they contain complex and intricate details.

Blurry Outputs: More often, autoencoder architectures lose fine details during the compression process. This results in blurry images with reduced quality. Often the images look blurry and with low sharpness [7].

Diffusion-Based Models

Diffusion models are another category of generative models that are often applied in text to image generation. Diffusion models introduce noise in a latent space and then remove it gradually to generate a high-quality output. Traditional diffusion models are computationally intensive and time-consuming. Popular models such as DALL-E-2 and GLIDE are diffusion models. They also consume high memory. Stable diffusion models are an improved version of traditional diffusion systems. These systems use latent space processing that makes them more efficient. They are faster than traditional diffusion models, computationally more efficient and occupy less memory. Leonardo.AI, Stable Video Diffusion and Hugging Face Diffusers are examples for systems using Stable Diffusion models. The following are the challenges identified in diffusion models and stable diffusion models.

Resource-heavy: Traditional diffusion models need extensive computational power and other resources for both training and inference. Especially, the iterative denoising phase requires a huge amount of computational power. However, some of the pre-trained models like stable diffusion models use low computer resources [8].

Longer Inference Times: Another challenge identified with diffusion models is that they are highly time-consuming. They require more time not only for training but also for inference [9].

Limited Applicability in Real-time Scenarios: Both high computational requirements and time constraints such models are found to be not appropriate for real-time applications in various domains and especially in resource-constrained devices [10].

Complexity in Hyperparameter Tuning: Hyper-parameter tuning in diffusion models is highly challenging because even minute changes in parameters can affect the quality of the output to a large extent and may result in low-resolution images. Such challenges make such models increase the complexity of adapting them in real-time application [11] .

Transformer-Based Models

Overview: Models like CLIP (Contrastive Language-Image Pretraining) have become foundational for understanding text-image alignment. These models are used in hybrid systems for text-to-image generation, such as DALL-E and OpenAI's

CLIP- based models. An extensive literature survey helped us identify the following drawbacks in transformer-based models.

Expensive Pretraining: Transformer-based models are data- hungry and require vast amounts of data and computational resources to train. This limits their accessibility to organizations or researchers without significant resources.

Coherence Issues in Complex Scenes: While good at generating simple images, transformer-based models can struggle with complex multi-object scenes, where intricate details and relationships between objects in the text description are important.

Bias and Generalization Problems: Transformer-based models are prone to inheriting biases from their training data, resulting in outputs that may lack diversity or reinforce harmful stereotypes especially when deployed in broader applications.

3. PROPOSED METHOD

The proposed method utilizes **Stable Diffusion**, a powerful text- to-image generation model, to create visually coherent storyboards from sequential text prompts. In the current approach, we have used descriptive prompts to generate individual storyboard frames and the individual frames are assembled to create a panel of images.

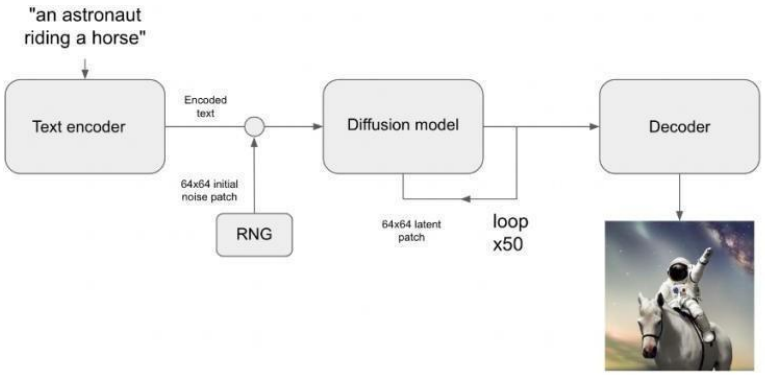

Fig. 1 Architecture of Stable Diffusion Model for Text to Image Generation.

The stable diffusion model is a diffusion-based deep learning model that is pre-trained on a vast dataset of images and captions. It generates images from the given prompt text. It interprets the descriptive prompts given by the user and generates an image that is aligned to the prompt text. A diffusion model creates images by using random noise in a latent space and gradually refining the model to produce high-quality image as output. The noises are iteratively refined until the desired image that semantically matches the prompt text is generated. Stable diffusion model's pretrained weights have been used to mitigate the need for extensive data preprocessing and model training. This reduces the computing requirements of the system and also makes faster inference. The architecture of the Stable Diffusion model used in the research work is shown in Fig. 1.

Fig. 2 Text to Image Generation.

The input is a prompt string that is passed through a text encoder. The text encoder encodes the input string into a numerical representation. The text encoder uses a Contrastive Language-Image Pretraining (CLIP) that is used for encoding text. A random noise generator (RNG) generates noise for the diffusion process. The diffusion model combines the encoded text

and noise pattern to generate a latent image representation. The model keeps on improving the image clarity at each step. The latent image representation is passed through a decoder which converts the latent representation into a full-resolution image that is the visual representation of the text prompt. The next section describes the phases of the text-to-image generation process.

Stable Diffusion Algorithm

The steps involved in the stable diffusion process is given below:

- Input: Text Prompt
- Encode the text prompt using CLIP text encoder
- Use a random blurry, noisy image (512x512) as input
- Represent the image in latent space (64x64) over 't' timesteps
- Forward Diffusion: Transform the image into pure noise
- Use a U-Net with cross-attention mechanism to align the generated image with the given prompt
- Reverse Diffusion: Progressively denoise the latent image using the U-Net neural network
- Convert the denoised latent representation into an image using VAE decoder
- Display the generated image

Text Encoding

The first crucial step of stable diffusion is text encoding. The model processes the input given by the user and interprets it to generate the image. The text encoder uses CLIP (Contrastive Language-Image Pretraining) to encode the prompt string. CLIP encodes the prompt string into a numerical representation to enable it to capture the meaning, context and the relationship between the different words in the prompt. The input text is first broken down into tokens using a tokenizer. The tokenizer breaks text into sub words or byte pair encodings to handle complex words. A numerical token ID is then assigned to each word or sub word. The tokenized text then passes through CLIP's text encoder and CLIP converts the token IDs into a high-dimensional vector representation based on the semantic meaning of the prompt. The final text embeddings are passed to the U-Net diffusion model.

Latent Noise Injection

The next phase of stable diffusion starts with random noise in the latent phase and denoising it gradually. The initial image is usually a Gaussian noise sample like a blurry image. The model performs the denoising in a lower dimensional latent space rather than working directly in the high-resolution pixel space. The original high dimensional image of the size 512 x 512 is compressed into a 64x64 latent space tenor by a Variational Autoencoder (VAE). The diffusion happens in two phases namely, forward diffusion and reverse diffusion. Real images are transformed into noise in the forward diffusion phase, and real images will be reconstructed from noise in the reverse diffusion phase. A deep neural network is used to remove the noise gradually in multiple iterations. Each step in the denoising process turns the random noisy, blurry image gradually into rough shapes and then into high-resolution images with detailed features. Since, the diffusion process happens in the latent space, the computational cost is reduced to a remarkable extent as compared to the traditional diffusion model. The VAE decodes the final latent representation back to the original high-resolution pixel space resulting in a high resolution image. The text embedding ensures that the final output aligns with the input prompt.

Prompt Conditioning

The prompt conditioning step is essential for guiding the model to create images that fit a narrative structure. A base prompt that is chosen initially would be the core theme of the storyboard (e.g., "A lady with an umbrella walking in rain"). Variations are appended to the base prompt to add different settings and this enables creating unique frames while maintaining a cohesive narrative. The core theme can be improved with different variations like, *'Lady walking in a road at sunset'* or *'Lady walking with a rainbow in the background'* or *'A girl walking in a beach'* which would create different variations of the basic picture suited to different contexts or environments. CLIP plays a significant role in prompt conditioning. CLIP generates text embeddings from the prompt that serve as conditioning inputs for the denoising U-Net model. CLIP prioritizes the most relevant words in the prompt e.g. rain, beach, sunset, rainbow etc. in this case. U-Net picks up the most relevant features for generating the final image.

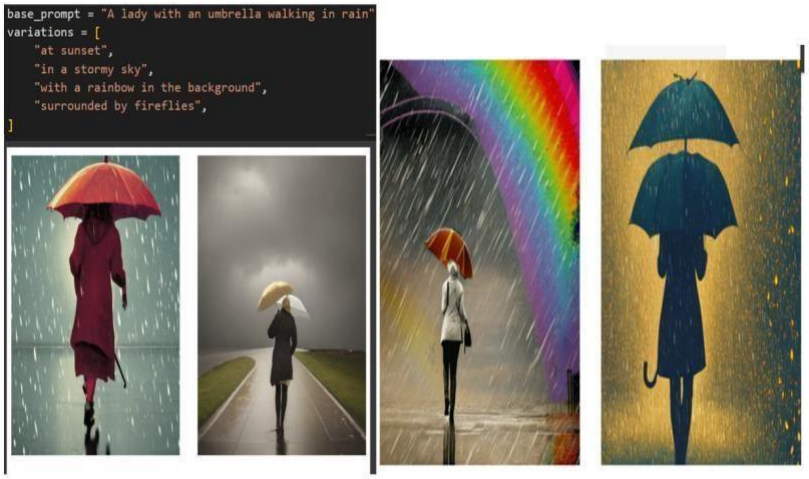

Fig. 3 Prompt Conditioning.

Fig. 3 shows the prompt conditioning for the core prompt '*A lady walking in rain with an umbrella'.* The core prompt represents the main theme of the image that would remain consistent in all environments. The context and environment may change through conditioning.

Storyboard Assembly

The diffusion models were also used to create storyboards or animations by creating a sequence of images. Thus, diffusion models can also be useful in applications that focus on animations, storytelling or comics where continuity is essential. The prompt conditioning can create individual frames with variations as prompted by the input string.

Fig. 4 Storyboard Assembly.

The individual frames can be combined in a sequence into a story panel, that would create a single image with a sequence. An image layout is chosen and frames are arranged in a sequence to convey the progression of the narrative visually. Python's Pillow library is used to merge images seamlessly into the storyboard layout, aligning each frame side-by-side creating a cohesive visual sequence. Customization options are available for the layout such as, vertical stacking of frames or a grid layout. Fig. 4 shows the assembling of a storyboard from a sequence of images.

4. RESULTS AND DISCUSSION

The stable diffusion model is implemented and the code was run on the NVIDIA RTX GPU. Pytorch, Hugging Face diffuser, and transformers were used for implementation. Contrastive Language-Image Pretraining (CLIP) was used to get the text embedding from the given input. A blurry and noisy random image has been chosen as the input and the image is represented using latent space representation in 64x64 dimension. A variational autoencoder was used to compress and reconstruct the images. A U-Net architecture featuring ResNet was used with Self-attention and Transformer layers. Cross-attention has been used to align image generation to the given input prompt. The model has been evaluated using different metrics namely, Frechet Inception Distance (FID), CLIP Score, Inception Score, Learned Perpetual Image Patch Similarity (LPIPS) and diversity score.

Frechet Inception Distance (FID) is a measure of the realistic nature of the generated images by comparing the feature distributions with the real image. The formula for FID is given in Equation (1).

$$FID = \| \mu_r - \mu_g \|^2 + Tr(\Sigma r + \Sigma g - 2(\Sigma r \Sigma g)^{1/2} \quad (1)$$

where,

μ_r, μ_g - mean feature vectors of real and generated images.

$\Sigma_r\Sigma_g$ - Covariance matrices of real and generated image features.

CLIP score evaluates how much the image aligns with the input prompt. The formula used for CLIP score is given in equation (2).

$$CLIP\ Score = \frac{E_i \,.\, E_t}{\|E_i\|\|E_t\|} \tag{2}$$

where E_i is CLIP image embedding and E_t is CLIP text embedding.

Inception score measures the quality and diversity of the image. It measures the quality and diversity of the images by analyzing their predicted class distributions using an Inception v3 model. LPIPS is used to measure the perceptual similarity. Intraclass variations are used to assess the diversity of the generated images. LPIPS calculates the difference in human perception rather than the pixel-wise differences. The formula for LPIPS is given in equation (3).

$$LPIPS(I_1, I_2) = \sum_l w_l \,.\, \|F_1(I_1) - F_1(I_2)\|_2^2 \tag{3}$$

The following table shows the various scores of our system.

Table 1. Evaluation Metrics

Metrics	Score
Frechet Inception Distance	12.7
CLIP Score	0.29
Inception Score	27.4
LPIPS	0.20
Diversity Score	0.77

In addition to the system generated metric, feedback has been obtained from around 50 users on realism, semantic accuracy and visual appeal of the images on a scale of 5. Users evaluated how realistic and natural looking the images are using the realism metric. The alignment of the image with the given prompt is measured using semantic accuracy. Users also rated the artistic quality of the images, color harmony, composition and aesthetic beauty of randomly chosen images. Each user was given 10 randomly selected images and they were rated. The average scores are given in Table 2.

Table 2 User Feedback

Image ID	Realism	Semantic Accuracy	Visual Appeal
Image 1	3.8	4.5	4.3
Image 2	4.2	4.2	4.2

Image 3	4.8	4.4	3.6
Image 4	4.9	3.5	3.8
Image 5	3.5	3.5	4.2
Image 6	4.7	4.2	4.7
Image 7	3.4	4.0	4.2
Image 8	4.5	4.5	4.3
Image 9	4.5	4.4	3.9
Image 10	4.2	4.3	4.5

The metrics indicate that the model performs fairly well in generating high quality images. Most of the images were visually appealing, with strong text alignment. Artistic images were produced in diverse styles. The inference time and computing complexity were also largely reduced. The following challenges were also faced. The model could not maintain consistency for complex images and multi-scenario scenes. Intricate patterns could not be represented accurately. The model was compared with existing models such as DALL-E and Imagen. The model could perform comparatively well in generating simple images. For complex scenarios, it was observed that the model requires improvement.

5. CONCLUSION

The research work analyzed the implementation of a stable diffusion model for generating sequential images and storyboard assembly with prompt conditioning. It was observed that the proposed model performed well in generating artistic and stylistic images. However, the model struggled to generate images related to complex prompts and also images related to scientific themes. The model can be applicable in various domains such as, entertainment, healthcare, e-commerce, education, social media and gaming. Future scope includes refining the model to generate images related to complex scenarios.

REFERENCES

[1] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas, StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks in Proceedings of IEEE International Conference on Computer Vision, ICCV 2017.

[2] Sawant, Ronit and Shaikh, Asadullah and Sabat, Sunil and Bhole, Varsha, Text to Image Generation using GAN (July 8, 2021). Proceedings of the International Conference on IoT Based Control Networks & Intelligent Systems - ICICNIS 2021, http://dx.doi.org/10.2139/ssrn.3882570

[3] Ramzan, S., Iqbal, M. M., & Kalsum, T. (2022). Text-to-Image Generation Using Deep Learning. Engineering Proceedings, 20(1), 16. https://doi.org/10.3390/engproc2022020016

[4] Rao, Abhishek & Bhandarkar, P & Devanand, Padmashali & Shankar, Pratheek & Shanti, Srinivas & Pai B H, Karthik. (2023). Text to Photo-Realistic Image Synthesis using Generative Adversarial Networks. 1-6. 10.1109/INCOFT60753.2023.10425482.

[5] L. Indira, M. Sunil, M. Vamshidhar, Ravi Teja, R. V. Praneeth. (2023). Text to Image Generation using GAN, International Research Journal of Engineering and Technology, 10(5), pp. 1479-1484.

[6] Gao, X., Fu, Y., Jiang, X., Wu, F., Zhang, Y., Fu, T., Li, C., & Pei, J. (2025). RSVQ-Diffusion Model for Text-to-Remote-Sensing Image Generation. Applied Sciences, 15(3), 1121. https://doi.org/10.3390/app15031121

[7] Mingzhen Sun, Weining Wang, Xinxin Zhu, Jing Liu. (2024). Reparameterizing and dynamically quantizing image features for image generation. Pattern Recognition. 146, https://doi.org/10.1016/j.patcog.2023.109962

[8] A. Rauniyar, A. Raj, A. Kumar, A. K. Kandu, A. Singh and A. Gupta, "Text to Image Generator with Latent Diffusion Models," 2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN), Ghaziabad, India, 2023, pp. 144-148, doi: 10.1109/CICTN57981.2023.10140348.

[9] Huan Li, Feng Xu, Zheng Lin, ET-DM. (2023). Text to image via diffusion model with efficient Transformer, Displays, 80, https://doi.org/10.1016/j.displa.2023.102568

[10] X. Hu et al., "Diffusion Model for Image Generation - A Survey," 2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR), Tianjin, China, 2023, pp. 416-424, doi: 10.1109/AIHCIR61661.2023.00073. keywords: {Surveys;Human computer interaction;Image synthesis;Superresolution;Robots;Research and development;Generative AI;AIGC;Diffusion Model;image generation;diffusion application},

[11] Sebaq, A., ElHelw, M. RSDiff: remote sensing image generation from text using diffusion model. (2024). Neural Comput & Applic 36, 23103–23111 (2024). https://doi.org/10.1007/s00521-024-10363-3

[12] Renato Sortino, Simone Palazzo, Francesco Rundo, Concetto Spampinato. (2023). Transformer-based image generation from scene graphs, Computer Vision and Image Understanding, 233, https://doi.org/10.1016/j.cviu.2023.103721

[13] Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., & Guo, B. (2022). StyleSwin: Transformer-based GAN for high-resolution image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11304–11314).

[14] S. R. Dubey and S. K. Singh, "Transformer-Based Generative Adversarial Networks in Computer Vision: A Comprehensive Survey," in IEEE Transactions on Artificial Intelligence, vol. 5, no. 10, pp. 4851-4867, Oct. 2024, doi: 10.1109/TAI.2024.3404910.

[15] S. Naveen, M. S. S Ram Kiran, M. Indupriya, T.V. Manikanta, P.V. Sudeep. (2021). Transformer models for enhancing AttnGAN based text to image generation, Image and Vision Computing, 115, https://doi.org/10.1016/j.imavis.2021.104284