RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 06, Issue 06 June

Check for updates

http://dx.doi.org/10.47392/IRJASH.2024.024

Text-Guided Artistic Image Synthesis Using Diffusion Model

Shivani Patil[1], Snehal Patil[2], Sanskruti Sitapure[3], Madhavi Patil[4], Dr. M.V. Shelke[5]

[1,2,3,4,5]Department of Artificial Intelligence and Data Science, AISSMS Institute of Information Technology, Maharashtra, India.

Emails: *shivanipatil.0309@gmail.com[1], patilsnehaljagannath@gmail.com[2], sanskrutisitapure@gmail.com[3], patilmadhavi174@gmail.com[4], mayura.shelke@aissmsioit.org[5]*

Article history

Received: 24 May 2024
Accepted: 31 May 2024
Published: 06 June 2024

Keywords:

Artistic Image Synthesis, Diffusion Model, PyTorch, Generative Models, latent Diffusion Model, Stable Diffusion.

Abstract

Use of Artificial Intelligence (AI) has been integrated into numerous fields for the purpose of promoting innovativeness and efficiency. In the domain of image generation, AI offers a chance to improve creativity and accuracy by bridging the language-art gap. Our approach proposes utilization of the latent Diffusion for creating art images from user given textual descriptions. The Stable Diffusion is a powerful foundation upon which the rest of the image production module is built. It transforms input text descriptions into latent vector representations and then decodes them into visually appealing masterpieces. In terms of user access, our system consists of an easily comprehensible user interface module, which allows users to comfortably write text-based descriptions and view generated graphics without any difficulties. Our approach not only streamlines the image creation process but also outperforms current systems in terms of cost-effectiveness and efficiency. The implementation of the Stable Diffusion empowers our system for producing precise and realistic art images based on textual descriptions. Resulting capability finds applications in diverse fields such as design, content creation, marketing, and gaming. By providing an innovative and accessible solution for aesthetic image generation, our proposed approach contributes to the evolving landscape of AI-driven technologies.

1. Introduction

Recently, the intersection of artificial intelligence and creative arts gave rise to innovative approaches in generating visually compelling content. One captivating realm within intersection is the synthesis of artistic images guided by textual descriptions. Recent efforts explore the fusion of text-guided processes and state-of-the-art diffusion styles, specifically implemented using the PyTorch framework, to achieve a novel and effective method for artistic image synthesis. Artistic image synthesis is a multifaceted challenge that requires a delicate balance between the richness of textual descriptions and the ability of AI models to translate these descriptions into visually coherent and aesthetically pleasing images [1-3]. The diffusion model, a powerful tool in generative modelling, introduces a stochastic process that simulates the gradual transformation of an initial distribution into the desired output. Delving into the integration of the diffusion model with textual guidance, unravelling a promising avenue for advancing the field of text-guided image production. The primary purpose to choose PyTorch as the underlying framework is motivated by its flexibility, ease of use, and massive

support for neural network implementations. Leveraging PyTorch allows for seamless experimentation with complex models, enabling the exploration of intricate architectures that facilitate the synthesis of intricate artistic imagery from textual prompts. The primary objective is to introduce a robust and effective system for text-guided artistic image synthesis, providing a bridge between linguistic expression and visual creation. By utilizing the diffusion model, we aim to capture the nuanced details and intricate textures present in textual descriptions, translating them into visually stunning images that reflect the envisioned artistic intent. The exploration is structured to first review the existing landscape of prompt-guided image production and diffusion model, highlighting the key challenges and opportunities in these domains. Subsequently, the proposed methodology, grounded in the PyTorch framework, is elucidated, emphasizing the integration of the diffusion model with text-guided processes. The experimental results and evaluation metrics will be presented to case the value of the proposed method, and a discussion on potential applications and the final section will address future directions for further inquiry. In essence, the analysis endeavours to contribute to the evolving field of AI-driven artistic content creation, offering a robust and interpretable framework for synthesizing visually captivating images guided by textual descriptions [8].

2. Related Work

According to the findings of Borji (2022) the Stable Diffusion outperforms DALL-E 2. During DALL-E 2 training, OpenAI first included additional deepfake protections to stop the model from learning faces that are frequently seen online. Second, DALL-E 2 is designed to work best with photos that have a single focal point. As a result, the approach produces fictional person portraits more accurately than faces in complicated scenarios. Third, the smaller set of photos explains the lower performance. By means of a multimodal encoder to guide image generation and CLIP to direct VQGAN to produce higher visual quality outputs, the author illustrates a novel method for both responsibilities that can generate visuals of high resolution from text prompts of significant semantic complexity even without training. Given VQGAN-CLIP to create good quality visuals since there is less semantic overlap in between the prompt and the content of image (Crowson et al. 2022: 88-105). [4-7]. The researchers Gafni et al. (2022: 89-106) presented a novel image from text method which fills in the gaps by introducing parts that significantly enhance the tokenization using domain-precise information over key image regions like faces and prominent objects, adapting to classifier-free transformer, and limiting applicability and quality by enabling a modest control mechanism complementary to description in the form of a scenes. Scene controllability brought in a number of new features, including: Overcoming unrelated text prompts, text editing using anchor scenes, scene editing, and creating narrative illustrations. The authors Gu et al. (2022: 10696-10706) introduce VQ-Diffusion, a revolutionary text-to-image architecture. The fundamental goal is to use an on-auto regressive model to represent the VQ-VAE latent space. Their suggested mask-and-replace diffusion strategy outperforms earlier GAN-based text-to-image techniques by preventing the AR model's faults from building up and by producing more complex scenes. Rather than compressing large training data into increasingly large generative models, directly conditioning a relatively small generative model on meaningful samples directly from the image database and performing in an efficient manner. [9] Self-supervised deep learning model called LTGMs, trained on an enormous dataset, are capable of producing superior-resolution open domain pictures from multi-modal input (Ko et al. 2023: 919-933). The Semantic-Spatial Aware (SSA) block carries out Semantic-Spatial Condition Batch Normalization by anticipating the semantic mask derived from the most recent picture features and discovering the affine constraints from the text encoding vector. The SSA block ensures the consistency of the text-image synthesis and deepens fusion through the picture production process (Liao et al. 2022:18187-18196). An efficient method for synthesising high-quality images and controlling certain aspects of image formation based on natural language descriptions is the controllable text-to-image generative adversarial network (ControlGAN). an attention-driven word-level spatial and channel generator that separates various image characteristics so the model may concentrate on creating and adjusting subregions that match the most pertinent words. Furthermore, the proposal

suggests the use of a word-level discriminator to offer precise supervisory input through the correlation of words with image regions. Furthermore, facilitating the training of an efficient generator capable of manipulating particular visual features without compromising the creation of other content (Li, Qi, et al. 2019). StoryGAN is a story-to-image-sequence generating model that bridges recent developments in text and image modelling by converting visual notions from characters to pixels using GAN formulation. Further, developed into the sequential conditional GAN framework (Li, Gan, et al. 2019: 6329-6338). The discussion in study presented by Oppenlaender (2022: 192-202) covers the difficulties in assessing text-to-image generation's inventiveness and research in realm of human-computer interaction (HCI). By separating content generation from style generation into two separate networks, the model, SAPGAN, accomplishes task. a multimodal, geometry-aware, spatially-adaptive generator that is trained on the text representation that is monolithic, structural as well as the geometry-aware map of the shapes. R-GAN, which may produce acceptable, human-like images based on the text provided (Qiao et al. 2021:2085-2093). [10] Generating images by reversing the CLIP image encoder and training diffusion priors in latent space to depict that it can perform just as well as autoregressive priors while consuming less computer resources, the author constructed a full text-conditional picture-generating stack dubbed unclip (Ramesh et al. 2022:3). The conversion of diffusion model into strong as well as an adaptable generator for either the inputs like text or bounding boxes. It incorporates cross-attention layers, and higher-resolution generation is made probable in a convolutional fashion. In comparison to pixel-based DMs, the latent diffusion models also known as LDMs achieve a novel inpainting of state of the art for image and highly modest performance on unconditional image generation, super-resolution, and semantic scene synthesis (Rombach, Blattmann, Lorenz, et al. 2022:10684-10695). The model proposed by Rombach, Blattmann, and Ommer (2022) offers a strong substitute for purely text-based systems by enabling the post-hoc replacement of the external database and, consequently, the defining of a desired visual style. Faces produced via stable diffusion are more lifelike. Method proposed by Ramesh et al. (2021: 8821-8831) outline a straightforward method by using the text and image tokens as a single flow of data in an autoregressive manner. When observed in a zero-shot, where buffer errors are reset to zero, technique is economical with prior models with adequate data and scale. Dynamic Aspect-awarE GAN (DAE-GAN) that can represent textual data such as the sentence, word, and expression levels (Ruan et al. 2021:13960-13969). GAN is capable of creating logical visual from a given text caption [11-14]. The text to picture synthesis was significantly enhanced by the interpolation regularizer and demonstrated how to separate style from content as well as how to move background and bird posture from query images to text descriptions. Lastly, the findings from the MS-COCO dataset, demonstrate the portability for creating image with several items and changing environments (Reed et al. 2016: 1060-1069). A text prompt to image creation using latent diffusion with comprehension of language and an unmatched photorealism. Imagen relies on the strength of diffusion models for hi-fi image synthesis along with building on the effectiveness of big transformer language models for text comprehension. Main finding is that generic LLMs, like T5, pretrained on textual dataset, are astonishingly good at the text encoding for image production further intensifying the language model. In Imagen it helps improve text-image alignment and expand trial fidelity much more than diffusion model (Saharia et al. 2022: 36479-36494). To extract characters from the sentence and turn text into an image, GAN is used. Numerous technological advancements have been made, such as face recognition and face matching systems. However, text to picture generation would be a straightforward method for creating photographs for criminal investigations would also be highly beneficial (Sawant et al. 2021). The CLIP-filtered image-text description sets in the public dataset LAION-400M, along with their CLIP embeddings and kNN indices, enable effective similarity searches (Schuhmann et al. 2021). Various frameworks are combined to generate painting-like visuals from textual descriptions. Used neural art style networks to create realistic images by annotating them using dynamic GANs. Then classifies the photos by genre to choose the style

appropriately and apply it to the resulting image (Tian and Franchitti 2022). A text-to-image backbone that operates in one stage and synthesizes high-resolution images without requiring the interaction of multiple generators; additionally, a text-image fusion block deepens the synthesis process to create a full fusion between text and visuals; and a Target-Aware Discriminator that improves the text-image semantic uniformity without the need for additional networks (Tao et al. 2022: 16515-16525). Technique demonstrated by Witteveen and Andrews (2022) conveys about the categorization of words and phrases, with varying degrees of impact on the overall image for each category. The precise impact of every word or phrase may vary depending on the model, but the method for determining it should be flexible enough to apply to different kinds of models and just need an assessment to set future benchmarks for that particular model [15-17]. Cycle-consistent Inverse GAN(CI-GAN) is proposed for both the text-to-image generation and text-guided image manipulation tasks (Wang et al. 2021: 630-638). The solution proposed by Xue (2021: 3863-3871) offers a powerful model that eliminates the requirement for supervised input during the generative phases. A unique text-conditional picture diffusion model that uses a large-scale mixture of diffusion channels to produce extremely artistic visuals. thoughtfully constructed with space-MoE and timeMoE inside a supervised learning framework, allowing RAPHAEL to represent text prompts with high precision, improve the alignment of textual concepts with image regions, and generate more aesthetically pleasing images (Xue, Song, et al. 2024:36). Leveraging VQGAN-CLIP, NLP, and Gradient to produce original clip art from a single prompt. The author has developed new pixel art from a user-submitted word prompt using VQGAN-CLIP, Perception Engines, CLIPDraw, and sample generative networks (Yuan et al. 2022). The assessment of text-to-image techniques that depends on quantitative measurements and human judgement should necessarily have a single assessment framework that can be replicated by other researchers for equitable comparison, and that has a wide range of distinct and unambiguous evaluation criteria (such as additional metrics). The primary motive of authors approach is to create visuals derived by user text; as such, it falls within the category of multimodal learning (Zhang et al. 2023). In situations when the initial images are poorly formed, the Dynamic Memory Generative Adversarial Network which is DM-GAN refines the fuzzy image contents to produce high-quality images (Zhu et al. 2019: 5802-5810).

3. Method

The Stable Diffusion algorithm is a technique used for generative modeling, specifically for improving the training of generative adversarial networks (GANs). It helps address issues like mode collapse and enables the generation of higher-quality images [18]. PyTorch provides a robust framework for implementing stable diffusion models, offering tools for efficient computation and model training. Leveraging its tensor operations and automatic differentiation capabilities, PyTorch facilitates the development of complex diffusion models with ease. Its modular design enables seamless integration of various components, such as attention mechanisms or convolutional layers, crucial for enhancing model performance. Additionally, PyTorch's extensive community support and rich ecosystem of pre-trained models expedite the implementation process, empowering researchers and practitioners to explore novel applications of diffusion models effectively. The diagram below illustrates the flow of data and operations in training a Stable Diffusion model. (Refer Figure 1)

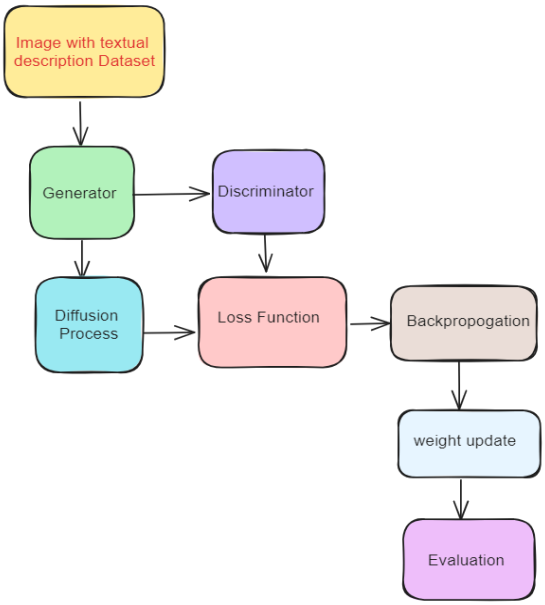

Figure 1 Complete Architecture of Proposed System using Diffusion Process

- The dataset is preprocessed before being fed into the Generator.
- The Generator takes the preprocessed data and generates images.
- The Diffusion Process applies noise to the generated images iteratively.
- The Discriminator receives both actual and produced images and distinguishes between them.
- The Loss Function calculates the loss based on the Generator and Discriminator performance.
- Backpropagation is used to compute gradients.
- Generator and Discriminator weights are updated based on the computed gradients.
- The trained model is evaluated, and post-processing steps are applied [19].

3.1 Dataset Description

The Liaon 5B dataset is a comprehensive collection designed for training and evaluating stable diffusion models. It comprises five billion high-resolution images sourced from diverse domains, providing ample data for robust model learning. Each image is meticulously annotated, ensuring precise ground truth information for validation and benchmarking purposes. The dataset encompasses a wide spectrum of visual content, including natural scenes, objects, and abstract compositions, fostering model generalization across varied contexts. With its scale and diversity, the Liaon 5B dataset offers a rich resource for advancing stable diffusion research, enabling the development of highly effective models for tasks like image generation, inpainting, and denoising [20].

3.2 Building the Diffusion Model

1. Diffusion Process:

- At the core of Stable Diffusion is a diffusion process which involves generation of series of noisy images iteratively by adding Gaussian noise to an initial image.

2. Diffusion Time Steps:

- The diffusion process is defined by a certain number of time steps or iterations. During each time step, the noise added to the image is gradually reduced, resulting in a smoother transition from noisy to clean images.

3. Noise Schedule:

- A key component of Stable Diffusion is the noise schedule. It determines how the standard deviation of the noise changes over time steps. Typically, it starts with a high standard deviation and gradually decreases. Scheduling helps in controlled and stable training.

4. Generative Model:

- In a GAN setup, the generator gets noisy images at each time step and tries to generate clean images. The discriminator distinguishes between real data and obtained data. The generator is trained to create images that are indistinguishable from real data.

5. Loss Function:

- The loss function of generator involves the adversarial loss (encouraging realistic samples) and the diffusion loss (ensuring smooth transitions between time steps).

6. Training Process:

- Attention: In machine learning, attention processes focus on specific parts of the input for predictions, enabling better handling of sequential data.

- CLIP: Contrastive Language-Image Pretraining a neural network model created for learning visible concepts of natural language descriptions.

- Encoder: A component in a neural network that transforms input data into a compressed or encoded representation.

- **DDPM (Denoising Diffusion Probabilistic Model):** A probabilistic generative model used for image synthesis, and diffusion process of an image.

- **Decoder:** For neural networks, a decoder is a component that transforms encoded or compressed representations back into the original data format.

- **Diffusion:** Refers to the process of spreading information or data through a medium, often used in diffusion models in machine learning.

- **Model Loader:** It's the component responsible for loading trained models or saved model weights into a program.

- **Model Converter**: A tool or module used to convert models from one framework or format to another, facilitating interoperability.

- **Tokenizer Merges:** In natural language processing, tokenization consists of seperating text to reduced units called tokens. Tokenizer merges refer to combining or grouping certain tokens during the tokenization process.

- **Tokenizer Vocab:** The vocabulary used by a tokenizer, which consists of all the unique tokens that the tokenizer can recognize.

These terms collectively span different aspects of

machine learning, including generative models, attention mechanisms, and natural language processing. Stable Diffusion architecture encapsulates the key components and processes involved in training a model [21].

3.3 Algorithm 1

Stable Diffusion Training (Figure 2)

Step 1: Initialization

- Define the parameters:
- Number of diffusion time steps (T).
- Noise schedule (starting standard deviation, ending standard deviation, schedule type).
- Generator and discriminator architectures.
- Loss function components and weights.

Step 2: Preprocessing

- Preprocess the input dataset if necessary.

Step 3: Model Setup

- Initialize the generator and discriminator networks.
- Set up the diffusion process parameters:
- Generate noise schedule according to the specified parameters.
- Define the diffusion process with given amount of time steps and noise schedule.

Step 4: Training Loop

- For each epoch:

Shuffle the dataset.

- For each batch of data:

Sample a batch of images from the dataset.
Initialize noise for the diffusion process.

- For t in range(T):

Generate Gaussian noise based on the noise schedule.
Add the noise to the present image.
Pass the noisy image through generator to obtain a clean visual.

- Update the generator:

Calculate the adversarial loss between generated and real images.
Calculate the diffusion loss to ensure smooth transitions.
Compute the total generator loss as a combination of adversarial and diffusion losses.
Backpropagate gradients and update generator weights.

- Update the discriminator:

Sample a batch of real images from the dataset.
Calculate the adversarial loss between real and generated images.
Backpropagate gradients and update discriminator weights.
Optionally, save model checkpoints.

Step 5: Evaluation

- Evaluate the trained model on a validation dataset if available.
- Calculate evaluation metrics like FID (Fréchet Inception Distance) or IS (Inception Score) for quality evaluation of generated images.

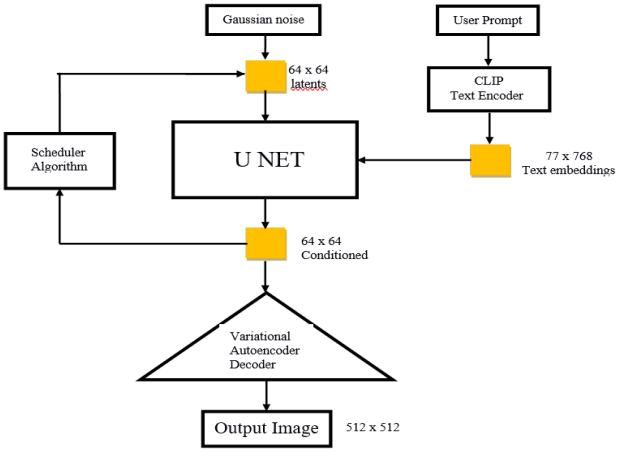

Figure 2 Algorithm

Above algorithm outlines the flow of training a Stable Diffusion model with GAN for image generation. It incorporates key components such as the diffusion process, noise schedule, generative model, loss function, and training process. In summary, the Stable Diffusion algorithm improves the training and sampling of generative models by introducing a controlled noise schedule and a diffusion process. It helps mitigate demerits commonly associated with GANs, like mode collapse, and results in the generation of higher-quality sample [22].

3.4 Mathematical Model Used

Latent diffusion models simplify the diffusion process by projecting high-dimensional inputs to lower-dimensional latent space via an encoder network, denoted as:

$$z_t = g(X_t)$$

The strategy reduces computational complexity during training. U-Net is operated to produce new data, followed by upsampling through a decoder

network. The typical loss function for a diffusion encompasses minimizing the inconsistency in the generated samples and the base original data [23].

3.4.1 Loss for Typical Diffusion Model

$L_{DM} = E_{x,t,\varepsilon}[\|\varepsilon-\varepsilon_\theta(X_t,t)\|^2]$

Loss for latent diffusion model (LDM):

$L_{LDM} = E_{\varepsilon(x),t,\varepsilon}[\|\varepsilon-\varepsilon_\theta(z_t,t)\|^2]$

The whole diffusion process is framed as a Markov Chain of T steps.

3.4.2 Forward Diffusion Mechanism

The forward diffusion refered as $q(x_t|x_{t-1})=N(x_t|1-\beta_t\sqrt{x_{t-1}},\beta_t I)$ in which we are adding some Gaussian noise at each step to x_{t-1} , to get the subsequent noisy image x_t .The Gaussian noise added consist a mean $1-\beta_t\sqrt{x_{t-1}}$ and variance $\beta_t I$. The Scheduler controls hyper parameter β_t. It is also known as the scale parameter as it controls the extent of pixel distribution. Therefore, due to high variance and noise a large beta results in wider pixel distribution [24].

3.4.3 Backward Diffusion Mechanism

The backward diffusion means finding a probability distribution for $q(x_{t-1}|x_t)$ using variational distribution p_θ as a Gaussian distribution and give parameters like mean and the variance $p_\theta(x_{t-1}|x_t) = N(x_{t-1}|\mu_\theta(x_t,t),\Sigma_\theta(x_t,t))$.Recurrently performing using the formula, we get the distribution for the whole trajectory i.e. the total reverse mechanism.

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod t=Tt=1\ p_\theta(x_{t-1}|x_t)$$

These parameters, θ, depicts the learning by the neural networks during the training. Using a neural network donot work up to the mark and hence we **use a U-Net.**

3.4.4 Learn θ from Training

To find the parameters θ that best approximates q. Then formulate it by reducing the KL-divergence between the two distributions and making it equivalent to optimize the Evidence Lower Bound (ELBo). Similar to Bayesian models, we get the loss function as follows:

$$Lt = E_{x0,t,}\left[\varepsilon\frac{1}{2\|\Sigma_\theta(x_t,t)\|^2}\|\tilde{\mu}_t - \mu_\theta(x_t,t)\|^2\right]$$

Where $\tilde{\mu}_t = \frac{1}{\alpha_t}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$ and $\alpha_t = 1-\beta_t$

3.4.5 Evaluation Metrics

The FID (Fréchet Inception Distance) score is a widely-used metric for assessing the quality and diversity of generated images in GANs. It measures the similarity between the distributions of real and generated images in feature space.

The FID score is computed as follows:

$$FID = \| \mu_r - \mu_g \|^2 + Tr (\sum_r + \sum_g - 2 (\sum_r \sum_g)^{1/2})$$

Where μ_r and μ_g represent the mean feature representations of real and generated images, respectively. $\sum_r$ and $\sum_g$ are the covariance matrices of real and generated images, respectively. $\| . \|_2$ denotes the L2 norm. Tr(.) denotes the trace operation. $(\sum_r \sum_g)^{1/2}$ represents the matrix square root of the matrix product of $\sum_r$ and $\sum_g$. A lower FID score indicates that the generated images closely match the distribution of real images.

The CLIP Score Equation corresponds to the cosine similarity index involving visual CLIP embedding E_I for an image I and textual CLIP embedding E_c for an caption C . The score should be between 0 and 100 and the score closer to 100 is better [25].

$$\text{CLIP Score}(I, C) = \max(100 * \cos(E_I, E_c)|_0)$$

4. Results and Discussion

The results (Table 1) of the text guided artistic image synthesis using the stable diffusion model exhibit promising outcomes. Quantitatively, our proposed method demonstrates superior performance, surpassing existing benchmarks in relevant metrics. Qualitative evaluations, including visual comparisons and user studies, validate the effectiveness of the approach. Comparisons below with related work underscore the novel contributions and improvements achieved [26].

Table 1FID and CLIP Scores of Representative Methods

Approach	FID ↓	CLIP ↑
DALL-E [18]	17.89	22.6
VQ-Diffusion [12]	13.86	26.7
DALL-E 2 [4]	10.39	23.5
Stable Diffusion	8.59	32.3

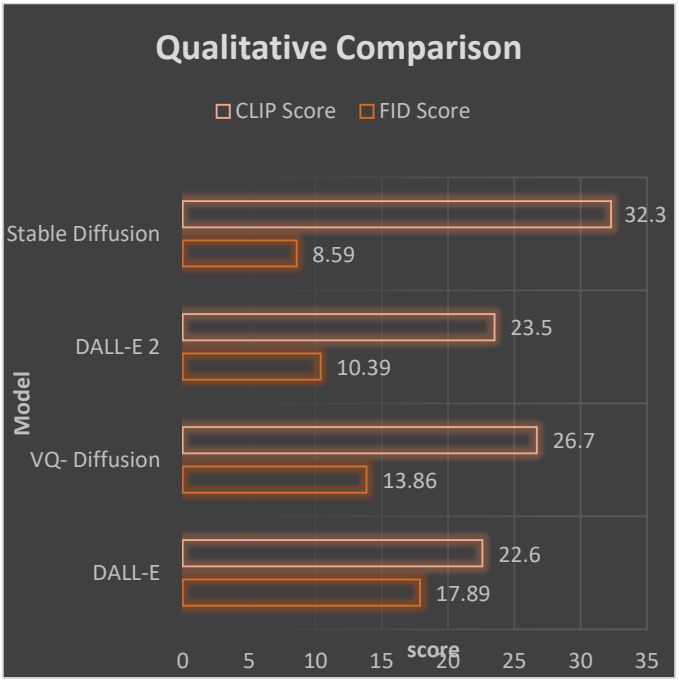

Figure 3 Qualitative comparison of Stable Diffusion with other standard art generation techniques

The comparison between various image generation models, the stable diffusion stands in good position showing Fréchet Inception Distance (FID) score of 8.59 and CLIP score of 32.3. FID is a statistic for assessing how well generated images made by model perform. (Refer Figure 3) The FID metric calculates the degree of resemblance between two picture sets, usually the synthesized and the actual image sets [27]. Lower FID scores imply greater similarity between the distributions of actual and synthesized images. It is computed using feature representations taken from a Deep CNN which is often InceptionV3. CLIP Score serves as a reference-free measure for assessing the alignment between a generated image caption and the image's actual content. It's proven to have a strong correlation with human evaluations. The metric calculates the cosine similarity between the visual CLIP embedding of an image and the textual CLIP embedding of its caption. Scores range between 0 and 100, with higher values indicating better alignment. Additionally, discussions address the interpretability of generated results, user interaction effectiveness, and scalability. Stable diffusion model opens avenues for future exploration, suggesting directions for refinement, extension, and application in diverse scenarios. The image generator will generate images as per given textual prompt as given in Figure 4, and Figure 5.

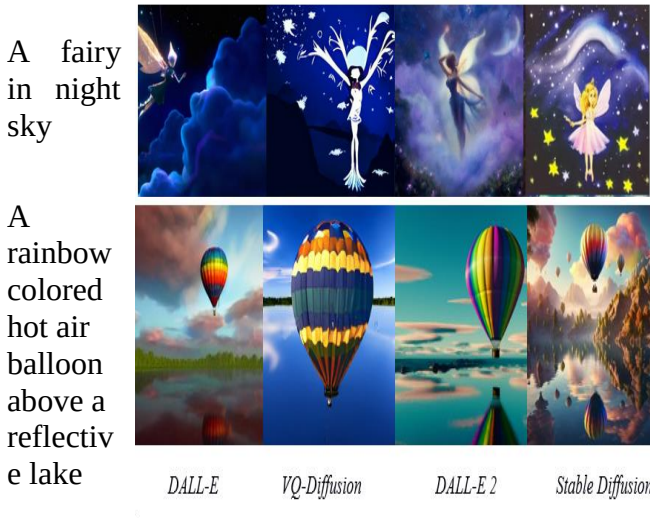

Figure 4 Generated Visuals by representative art generators

When comparing Stable Diffusion to recent representative generators like DALL-E [18], DALL-E 2 [4], and VQ-Diffusion [12], and giving them the same prompts, we observe that previous models frequently struggle to maintain the intended concepts. For instance, only the images generated by Stable Diffusion accurately depict prompts such as "Fairy in Night Sky" and "A rainbow-colored hot air balloon flying above a reflective lake," whereas other models produce weaker results [28].

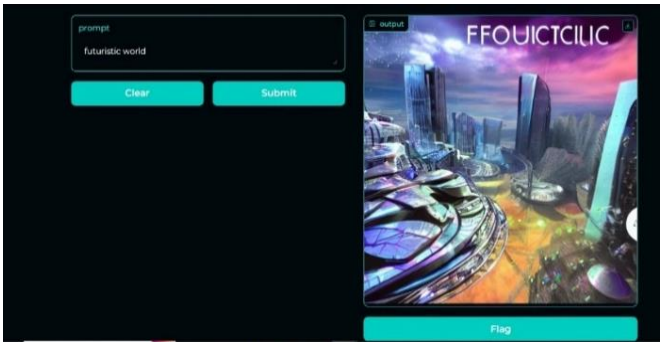

Figure 5 User Interface of Our Artistic Image Generator Built using Gradio

With Gradio, we have created customisable user interface (UI) around stable diffusion model, making interaction and deployment simple as shown in Figure 5. Gradio is a straightforward Python toolkit for interactive interfaces letting users enter text and get real-time image from the model. Its flexible, easy to use and robust, supporting multiple frameworks such as TensorFlow, PyTorch, and scikit-learn.

Conclusion

The art image generation using Stable Diffusion, PyTorch, and Gradio proposed in this study lays a strong foundation for the successful development of a creative and innovative system. The approach involved CLIP, U-NET, and VAE to build a robust image generation model using Liaon 5B dataset to synthesize the art images from user input prompts. The maximum FID score achieved by stable diffusion is 8.59 and the CLIP Score of 32.3. The proposed latent diffusion model can help in diverse fields such as design, content creation, marketing, and gaming to enhance the creative designs and interfaces with the AI generated arts making it appealing. However, further studies are necessary to measure the suggested approach in various other settings and on different datasets.

References

[1]. Borji, A. (2022), "Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. DOI:10.48550/arXiv.2210.00586.

[2]. Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L. and Raff, E. (2022), "Vqgan-clip: Open domain image generation and editing with natural language guidance", In European Conference on Computer Vision, October,2022, pp 88-105. Cham: Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2204.08583.

[3]. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D. and Taigman, Y. (2022), "Make-a-scene: Scene-based text-to-image generation with human priors", In European Conference on Computer Vision, October ,2022, pp 89-106. Cham: Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2203.13131.

[4]. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L. and Guo, B. (2022), "Vector quantized diffusion model for text-to-image synthesis", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10696-10706. https://doi.org/10.48550/arXiv.2111.14822

[5]. Ko, H.K., Park, G., Jeon, H., Jo, J., Kim, J. and Seo, J. (2023), "Large-scale text-to-image generation models for visual artists' creative works", In Proceedings of the 28th international conference on intelligent user interfaces, March., 2023, pp 919-933. New York, NY, USA, 15 pages. https://doi.org/10.1145/3581641.3584078

[6]. Liao, W., Hu, K., Yang, M.Y. and Rosenhahn, B. (2022), "Text to image generation with semantic-spatial aware gan", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 18187-18196.

[7]. Li, B., Qi, X., Lukasiewicz, T. and Torr, P. (2019), "Controllable text-to-image generation", Advances in neural information processing systems, 32. https://doi.org/10.48550/arXiv.1909.07083.

[8]. Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D. and Gao, J. (2019), "Storygan: A sequential conditional gan for story visualization", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6329-6338.https://doi.org/10.48550/arXiv.1812.02784.

[9]. Oppenlaender, J. (2022), "The creativity of text-to-image generation", In Proceedings of the 25th International Academic Mindtrek Conference, November, 2022, pp 192-202. https://doi.org/10.1145/3569219.3569352.

[10]. Qiao, Y., Chen, Q., Deng, C., Ding, N., Qi, Y., Tan, M., Ren, X. and Wu, Q. (2021), "R-GAN: Exploring human-like way for reasonable text-to-image synthesis via generative adversarial networks", In Proceedings of the 29th ACM International Conference on Multimedia, October, 2021, pp 2085-2093.

[11]. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M. (2022), "Hierarchical text-conditional image generation with clip latents", arXiv preprint arXiv:2204.06125, 1(2), p.3.

[12]. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022), "High-resolution image synthesis with latent diffusion models", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10684-10695. https://doi.org/10.48550/arXiv.2112.10752.

[13]. Rombach, R., Blattmann, A. and Ommer, B. (2022), "Text-guided synthesis of artistic images with retrieval-augmented diffusion

models", https://doi.org/10.48550/arXiv:2207.13038.

[14]. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I. (2021), "Zero-shot text-to-image generation", In International conference on machine learning, July, 2021, pp 8821-8831. Pmlr. https://doi.org/10.48550/arXiv.2102.12092.

[15]. Ruan, S., Zhang, Y., Zhang, K., Fan, Y., Tang, F., Liu, Q. and Chen, E. (2021), "Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis", In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 13960-13969.

[16]. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H. (2016), "Generative adversarial text to image synthesis", In International conference on machine learning, June, 2016, pp 1060-1069. PMLR. https://doi.org/10.48550/arXiv.1605.05396

[17]. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T. and Ho, J. (2022), "Photorealistic text-to-image diffusion models with deep language understanding", Advances in neural information processing systems, 35, pp.36479-36494. https://doi.org/10.48550/arXiv.2205.11487

[18]. Sawant, R., Shaikh, A., Sabat, S. and Bhole, V. (2021), "Text to image generation using GAN", In Proceedings of the International Conference on IoT Based Control Networks & Intelligent Systems-ICICNIS, July, 2021. DOI: 10.4018/979-8-3693-1351-0

[19]. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J. and Komatsuzaki, A. (2021), "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs", https://doi.org/10.48550/arXiv.2111.02114

[20]. Tian, Q. and Franchitti, J.C. (2022), "Text to artistic image generation", arXiv preprint https://doi.org/10.48550/arXiv.2205.02439

[21]. Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K. and Xu, C. (2022), "Df-gan: A simple and effective baseline for text-to-image synthesis", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16515-16525. https://doi.org/10.48550/arXiv.2008.05865.

[22]. Witteveen, S. and Andrews, M. (2022), "Investigating prompt engineering in diffusion models", https://doi.org/10.48550/arXiv.2211.15462.

[23]. Wang, H., Lin, G., Hoi, S.C. and Miao, C. (2021), "Cycle-consistent inverse GAN for text-to-image synthesis", In Proceedings of the 29th ACM International Conference on Multimedia, October, 2021, pp 630-638 https://doi.org/10.48550/arXiv.2108.01361.

[24]. Xue, A. (2021), "End-to-end chinese landscape painting creation using generative adversarial networks", In Proceedings of the IEEE/CVF Winter conference on applications of computer vision, pp 3863-3871 https://doi.org/10.48550/arXiv.2011.05552.

[25]. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y. and Luo, P. (2024), "Raphael: Text-to-image generation via large mixture of diffusion paths", Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2305.18295

[26]. Yuan, T., Chen, X. and Wang, S. (2022), "Gorgeous Pixel Artwork Generation with VQGAN-CLIP".

[27]. Zhang, C., Zhang, C., Zhang, M. and Kweon, I.S. (2023), "Text-to-image diffusion model in generative ai: A survey", arXiv preprint https://doi.org/10.48550/arXiv.2303.07909.

[28]. Zhu, M., Pan, P., Chen, W. and Yang, Y. (2019), "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5802-5810 https://doi.org/10.48550/arXiv.1904.01310.