Date of publication DEC-31, 2024, date of current version AUG-28, 2024
www.computingonline.net/ computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.23.4.3766

Improving Text-Driven Image Synthesis: Diffusion Models for Photorealistic Outcomes

T.M.N.VAMSI[1], J. N.V.R. SWARUP KUMAR[1],I.S.SIVA RAO[2],PRATIBHA LANKA[3]
[1]Department of CSE, GST, GITAM Deemed to be University, Visakhapatnam, India. (e-mail: mthalata@gitam.edu, sjavvadi2@gitam.edu)
[2]Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India (e-mail: prof.sundarasivarao@andhrauniversity.edu.in)
[3]Department of Computer Science and Engineering, Engineering Technology Program, GVPCDPGC(A), Visakhapatnam, India (e-mail: pratibha@gvpcdpgc.edu.in)
Corresponding author: T.M.N.Vamsi (e-mail: mthalata@gitam.edu).

ABSTRACT In recent developments, there has been a noteworthy demonstration of the effectiveness of generating high-quality images of diffusion models. This success is further enhanced when these models are combined with a technique that allows for a strategic balance between image diversity and fidelity. Addressing the challenge of text-conditional image synthesis, we extensively explore the utility of diffusion models along with two distinct guiding approaches: CLIP (Contrastive Language–Image Pretraining) guidance and classifier-free guidance. Through a comprehensive analysis, we uncover intriguing insights. The classifier-free guidance method consistently emerges as a standout performer, producing images with remarkable photorealism. This method showed a PSNR of 183.66 dB and an SSIM of 99.99%, indicating efficient photorealism and structural similarity to ground reality images. It presents a unique approach that combines diffusion models with classifier-free guidance for text-conditional image synthesis, focusing on photorealism and alignment with captions. Therefore, it can be useful for human evaluators to proficiently maintain both visual realism and associated captions.

KEYWORDS Text-Conditional; Diffusion Models; Photorealistic Images; CLIP; Classifier-Free Guidance; GANs (Generative Adversarial Networks); Transformer-Based Neural Network; Latent Code Encoding; Inpainting; Perceptual Loss.

I. INTRODUCTION

IN the contemporary communication and visual expression landscape, images have transcended mere visual aids to become intricate vehicles for conveying complex messages and emotions. However, the creation of images has traditionally demanded a combination of artistic skill, technical prowess, and significant time investment. The advent of modern text-conditional image models has redefined this paradigm, offering the tantalizing prospect of generating images from textual descriptions. This innovation democratizes visual content creation and paves the way for a new era of seamless integration between language and imagery. Nevertheless, pursuing photorealism in such generated images has remained an elusive challenge. In response to this challenge, researchers have ventured into uncharted territories to imbue text-conditional image models with the coveted attribute of photorealism within a class-conditional framework. Two distinct pathways have emerged as beacons of progress in this pursuit. The work of Dhariwal Nichol (2021) [1] introduced a fusion of diffusion models and classifier guidance, leading to a symbiotic relationship between generative algorithms and pre-trained classifiers. In a parallel narrative, Ho Salimans (2022) [2] defied convention by proposing a classifier-free guidance approach, showcasing the potential for generating photorealistic images without explicit reliance on a separately trained classifier. This research propels innovation trajectory further, introducing a groundbreaking model redefining image synthesis boundaries. Distinguished by its autonomous capacity to produce photorealistic images without the crutch of a classifier, this model is a testament to the amalgamation of state-of-the-art technologies and sophisticated architectural design. The resulting images rival and surpass the photorealism of samples produced by DALL-E [3], a pioneering text-to-image synthesis model. Impressively, in 87% of comparisons, the proposed model's images are deemed more photorealistic, and in 69% of cases, they align more closely with accompanying captions. The journey to develop and refine this innovative model necessitates a robust foundation. Drawing from the vast expanse of the internet, an extensive dataset of text-image pairs was curated. Stringent curation protocols were enforced to curtail the inclusion of sensitive or harmful content. The resulting dataset, comprising a staggering 67 million text-image pairs, was seamlessly integrated with an established dataset that underpinned the training of CLIP models [4]. This fusion, culminating in an enriched dataset of approximately 137 million pairings, is the cornerstone of the model's training and underpins its remarkable achievements.The convergence of generative prowess, architectural ingenuity, and ethical consideration forms the bedrock upon which this investigation unfolds, offering a glimpse into the future of human-computer co-creation and redefining the boundaries of visual expression.

II. STATE OF THE ART

Image synthesis has garnered substantial attention in computer vision in recent years. Diverse generative models have emerged as formidable tools for crafting high-quality images, encompassing prominent techniques such as Generative Adversarial Networks (GANs) [5]. Variational Autoencoders (VAEs) is a stochastic variational inference method for efficiently handling large datasets with continuous latent variables and intractable posterior distributions [6]. Autoregressive models by A¨aron van den Oord et al. [7] present a deep neural network for modeling natural images by predicting pixel values sequentially, addressing expressiveness, tractability, and scalability. It achieves superior log-likelihood scores compared to prior methods, serving as a benchmark on ImageNet and generating high-quality, coherent image samples. However, these models bear inherent limitations when generating images with profound semantic meaning.

One of the early pioneers in this arena was Reed et al. [8], who harnessed the potential of GANs to give life to images from textual prompts. It is a novel deep architecture and GAN approach to synthesizing realistic images from text descriptions, bridging the gap between text and image modeling. The model effectively generates credible bird and flower images based on detailed textual descriptions. Since then, a panorama of text-to-image generation methodologies has unfolded, encompassing notable contributions such as Stacked Generative Adversarial Networks (StackGAN) to generate high-quality 256x256 images from text descriptions. The approach decomposes the problem into two stages, effectively refining primitive shapes and adding realistic details [9], Attentional Generative Adversarial Network (AttnGANan) [10] for fine-grained text-to-image generation. It uses attention mechanisms to refine image details based on relevant text words, achieving significant performance improvements over prior methods, and a novel framework called MirrorGAN [11] is presented for text-to-image generation that focuses on ensuring visual realism and semantic consistency. Another approach in which an innovative avenue through their blended diffusion is introduced by Avrahami et al. [12], which is a novel scheme for generating images from textual descriptions. Their methodology entails a diffusion process that constructs an initial image, subsequently refined through a meticulous image blending process. An analogous concept was pursued by Bau et al. [13] in their "Paint by Word" approach, empowering users to manipulate and edit images utilizing natural language descriptors seamlessly.

Recent strides in this domain have birthed the "clip-guided diffusion model," an ingenious proposition pioneered by Nalisnick et al. [14]. This approach excels in its computational efficiency vis-à-vis traditional GAN-based methods and proficiency in enabling meticulous control over the generated image through textual input. The pioneering work of Brock et al. [15] birthed a colossal GAN training methodology that substantially elevates the quality of generated images. This advancement has been integrated into several text-to-image generation frameworks, including the innovative StackGAN and MirrorGAN [9], [11] paradigms.

The research paper [16] investigates the application of large-scale language-image models, specifically text-guided diffusion models, for image editing. It highlights significant advancements in image generation, emphasizing the ability to create photorealistic images in various domains. The research paper [17] presents a novel approach to editing real images using text inputs utilizing advanced diffusion models. This method enables significant image alterations based solely on textual descriptions.

The paper [18] [19] comprehensively reviews the advancements in Generative Adversarial Networks, a pivotal concept in artificial intelligence and deep learning. It delves into the foundational principles of GANs, charts their evolution, and discusses various enhancements and applications, including image generation and data augmentation. As a survey, it synthesizes a broad spectrum of research and developments, providing insights into GAN technology's latest trends and potential future directions.

The paper [20] presents an innovative approach to generating realistic X-ray images for security purposes. The images are essential for training and evaluating security screening systems. This work significantly contributes to security imaging, offering a new tool for enhancing security systems and improving training protocols.

The paper by researchers from Rutgers University and Snap Inc. introduces "SINE," [21] a novel approach to editing single images using text-to-image diffusion models. The paper emphasizes leveraging large-scale pre-trained diffusion models for real image editing, a significant advancement in AI-driven image manipulation. SINE's methodology involves adapting text-to-image diffusion models to maintain resolution and quality while editing real images.

The research paper "Learning to See by Moving" [22] [23] by Agrawal, Carreira, and Malik proposes a novel approach to feature learning in computer vision, shifting away from the conventional reliance on hand-labeled images for training neural networks. Instead, it explores egomotion—awareness of one's own movement—as an alternative supervisory signal. This approach is inspired by biological organisms that develop visual perception primarily for navigation and interaction within their environment.

The research paper [24] by Lucic et al. presents a significant advancement in deep generative models, particularly focusing on conditional generative adversarial networks (GANs) for natural image synthesis. The work stands out for its ability to generate high-quality, diverse images at high resolution with a substantially reduced reliance on labeled data. By integrating self- and semi-supervised learning methods, the authors successfully challenge the conventional heavy dependence on vast labeled datasets in image generation. Their approach matches and surpasses the performance of the state-of-the-art BigGAN model on the ImageNet dataset, achieving this feat with only 10% to 20% of the labels typically required. This breakthrough signifies a major leap in inefficient and resource-effective image generation, marking a notable shift towards more accessible and scalable generative modeling in machine learning.

A. ADVANCEMENTS OF THE PROPOSED SYSTEM

The proposed system heralds a new era of advancement by fusing cutting-edge methods to transcend the constraints of existing paradigms. Where conventional models falter in semantically meaningful image generation, the proposed system navigates a unique trajectory guided by textual descriptions. Key differentiators lie in the innovative integration of the blended diffusion approach and the StackGAN architecture, ushering forth a model that excels in computational efficiency and granular control. Fusing these techniques empowers the system to transcend the limitations that have historically hindered the domain. By seamlessly converging the expressiveness of natural language with the intricacies of image generation, the proposed system emerges as a torchbearer of transformative potential. This model's profound capacity for precise image manipulation and seamless integration of textual cues engenders a paradigm shift in creative expression. The user's ability to orchestrate intricate visual narratives through language becomes not only a reality but an immersive experience. The proposed system's unique synthesis of techniques bridges the chasm between textual guidance and image synthesis, transcending the boundaries of prior methodologies. While prior works laid the groundwork, the proposed system erects a bridge that transcends the limitations of each individual approach. This synthesis results in a harmonious union, catapulting the realm of image synthesis into a new era characterized by unprecedented fidelity, efficiency, and expressive potential.

III. PROBLEM STATEMENT AND OBJECTIVES

A. PROBLEM STATEMENT

The challenge of generating photorealistic images from textual prompts poses a formidable obstacle at the intersection of natural language processing and computer vision. Traditional text-to-image models encounter shortcomings in their ability to produce images that are both coherent and true to life. This limitation becomes particularly pronounced when tackling complex and abstract ideas, hindering the seamless integration of language and visual representation.

B. OBJECTIVE

The primary objective of the Guided Language-Image Diffusion Enhancement (GLIDE) [25] model is to transcend the shortcomings entrenched within prevailing text-to-image models. This pursuit is achieved by strategically deploying a diffusion-based generative model designed to harness textual descriptions' complexities and transform them into an ensemble of high-quality, photorealistic images. Importantly, the GLIDE model embodies innovation in its capacity to generate images and its potential for interactive image refinement. By affording users the unique ability to modify and enhance generated images meticulously following their specific requisites, the GLIDE model pioneers an era of dynamic and collaborative image creation that amalgamates human ingenuity with computational prowess.

IV. PROPOSED METHODOLOGY AND ALGORITHM

This method describes the diffusion models to meticulously craft images that adhere to the conditions specified in the text and exude photorealistic qualities. The methodology central to this work bridges the gap between textual prompts and their corresponding visual representations through text-conditional image synthesis. At its core, the process employs diffusion models, recognized for their ability to propagate and integrate information, ensuring that the synthesized images accurately mirror the textual cues. Unlike many existing systems that leverage classifiers to guide the generation process, this approach stands out because of the adoption of a classifier-free guidance mechanism. This innovative direction ensures the images produced are coherent with the textual conditions and attain a photorealistic quality. This combined effect between textual understanding and diffusion modeling facilitates the creation of images that capture the essence of the described scene or object with remarkable accuracy. The entire methodology is described below in Figure 1 and Figure 2.

As depicted in Figure 1, the text encoding to image upsampling process consists of the following steps.

Step1: Encoding the Text: - Utilize a transformer-based neural network to convert the input text prompt into a latent representation, yielding a fixed-length latent code. Append this latent code to the initial noise generated for a 64x64 image resolution. Step2: The Input to the model: - Present the Multi-Scale Gradient Descent (MSGD) with two inputs: i. A low-resolution conditioning image. ii. A random noise

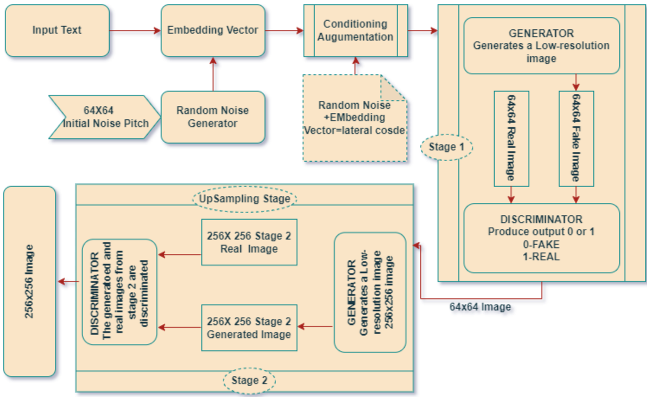


Figure 1. Text Encoding to Image Up sampling process.

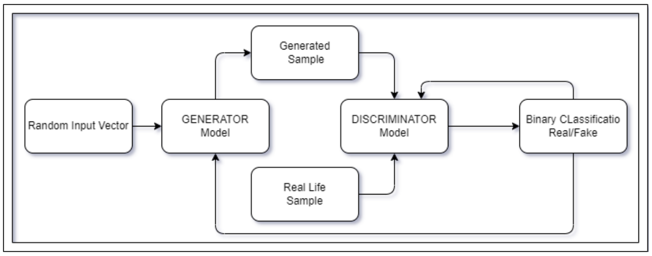


Figure 2. Generator-Discriminator block for Real vs. Fake Image Checking.

tensor, designated as the primary input for the generator network.

Step3: The Multi-Scale Diffusion (MSD) [26] Process: - MSD takes the input noise tensor through a series of diffusion steps. The tensor's data is spread out in this progression, retaining the foundational image structure. Starting from the broadest scale, MSD fine-tunes the image, moving towards more refined scales.

Step4: Conditioning at Various Scales: - During each diffusion stage, the generator network works with a mix of the diffused noise tensor and a feature map derived from the conditioning image at that scale. This combination ensures the generator produces detailed images that align with the provided reference.

Step5: Progressive UpSampling: - Post every diffusion step, upsample the output tensor to match the forthcoming scale. Repeat the diffusion and conditioning processes at this new scale. This iterative process empowers the generator network to enhance the image with finer details at each scale increment.

Step6: Final Output Generation: - By scaling up the last diffusion step's output tensor to its original dimensions, MSGD crafts its final, high-quality visual output.

As depicted in Figure 2, the explanation below describes checking whether the generated image in process 1 is real or fake. Also, in Figure 2, the neural network structure encompasses both a generator and a discriminator network. When presented with random noise as input, the generator network generates a data sample, subsequently fed into the discriminator network alongside actual data samples. Subsequently, the discriminator network endeavors to distinguish between genuine and spurious inputs. This process, termed adversarial training, involves the simultaneous refinement of both networks. While the discriminator network enhances its ability to distinguish authentic and counterfeit data, the generator network progressively improves its capacity to generate remarkably lifelike data samples, adept at deceiving the discriminator. Therefore, the approach explained above delves into the intricacies of this method, highlighting its potential and efficacy in generating lifelike synthesized images from textual prompts.

Algorithm 1 Text-to-Image Generation with Guided Diffusion

```
Require: Input_Text: Text prompt
Ensure: Generated_Image: Generated image
Latent_Code = TextEncoder(Input_Text)
for each scale $s$ do
    Add Gaussian Noise $\epsilon$ to Image_s
    Image_s = Image_s + $\epsilon$
    Image_s = Apply_Diffusion_Steps(Image_s)
    Reduced_Noise_Image_s = Conditional_Noise_Reduction(Image_s, Latent_Code)
    Denoised_Image_s = Denoise(Reduced_Noise_Image_s)
end for
Generated_Image = Synthesis_Network(Denoised_Image, Latent_Code)
Perceptual_Loss = Calculate_Perceptual_Loss(Generated_Image, Reference_Image)
Optimize(Synthesis_Network_Params) to minimize Perceptual_Loss
Modified_Latent_Code = Modify_Latent_Code(Latent_Code)
Edited_Image = Synthesis_Network(Denoised_Image, Modified_Latent_Code)
return Generated_Image
```

The algorithm follows a systematic flow to achieve its goal of generating photorealistic images from textual prompts. The process commences by encoding the provided textual prompt into a latent code using a text encoder network, typically a transformer-based model. This code serves as the foundation for subsequent image generation.

The core of the algorithm revolves around the multi-scale guided diffusion process. At each scale, the algorithm introduces controlled noise to the image, gradually diffusing it through a series of transformation steps while retaining high-level features. This is followed by a conditional noise reduction step, where noise reduction is guided by the latent code using an attention mechanism. The final denoising operation results in the desired image at that scale.

With the guided diffusion process completed, the synthesis network takes over. It combines the outcomes of the diffusion process and the latent code to generate a new image that aligns with the given text prompt. The generator model, a synthesis network, leverages the attributes of the latent code to produce a high-quality image.

Fine-tuning refines the generated image through a perceptual loss function, which measures its similarity to a reference image aligned with the text prompt. This enhances the accuracy and quality of the final output.

Finally, the algorithm accommodates image editing, allowing users to modify the latent code for attribute changes, style variations, or other adjustments. The synthesis network then regenerates the image according to the modified latent code, yielding an edited version of the original image.

V. RESULT ANALYSIS AND DISCUSSION

In our qualitative analysis, we observe that the images generated by GLIDE (filtered) often exhibit a partially realistic appearance. However, the model's relatively modest size constrains its ability to associate attributes with objects and accomplish complex compositional tasks. This limitation also affects GLIDE's breadth of knowledge, especially in contexts involving humans. This is because the dataset used for training GLIDE (filtered) underwent preprocessing to remove images of people. Additionally, compared to models of similar dimensions trained on our internal dataset, GLIDE encounters challenges in effectively combining multiple objects in intricate ways. This challenge arises from the specific characteristics of the dataset used for GLIDE's training (filtered). It's important to note that we haven't conducted direct quantitative testing for GLIDE (filtered). This is particularly noteworthy as many evaluation prompts used heavily involve generating images containing people, which could potentially introduce bias against GLIDE (filtered) in our reported evaluations. However, we acknowledge that there are various methods for assessing performance, and these will be comprehensively discussed in the results section. The results are obtained by submitting the text input to the algorithm, and the sampled output for a specific input is presented below.

A. ILLUSTRATION OF THE PROPOSED ALGORITHM THROUGH THE FOLLOWING EXAMPLE:

Prompt Description: "A peaceful river with a bridge."

Text Transformation: A specialized encoder translates the provided text into a latent representation.

Guided Diffusion Across Scales: Starting with a randomized noise pattern, guided diffusion incorporates the latent representation using a diffusion model. This process unfolds iteratively, traversing multiple scales that are progressively more refined. Each iteration involves adding Gaussian noise, diffusing this noise, selectively reducing noise in sections identified by the latent code, and culminating in a denoising step to finalize the image at that scale.

Image Creation: Drawing from the outcomes of the diffusion steps and the latent code, the synthesis mechanism, typically a generator network, crafts a detailed image in line with the text description. The result is a finely generated representation.

Image Refinement: The initially synthesized image is refined by gauging it against a benchmark image corresponding to the textual prompt, using a perceptual loss metric. The goal is to fine tune the network parameters to reduce discrepancies between the synthesized and reference images.

Adaptive Editing: Images can be edited precisely by tweaking the latent representation and reusing the synthesis network. For example, changing the river's state from peaceful to aggressive would involve modifying the latent code to reflect this change and regenerating the image, allowing various edits to the generated output.

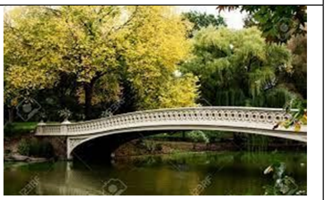
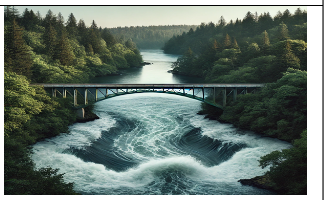
Input: Prompt	**Image generated by proposed Model**
A peaceful river with a bridge	
Input: Prompt	**Image generated by proposed Model**
An aggressive river with a bridge	

Table 1. Generated image based on the input prompt.

Analysis of Quality Parameters: A high PSNR score means less distortion and noise in the generated image, which closely resembles the reference image regarding pixel precision. This indicates that using the Guided Diffusion algorithm, the Text-to-Image Generation effectively eliminates differences to produce a crisp and detailed image. A high PSNR score indicates that the algorithm may produce quantitatively and visually attractive images appropriate for high-precision applications.

Measured against the reference image in brightness, contrast, and structural integrity, the high SSIM value provides additional quality assurance. It shows that the generated image faithfully captures the original image's structural characteristics and perceptual qualities. Also, it accurately represents the visual components, such as the trees, bridge, and scene arrangement. The algorithm's ability to preserve structural and perceptual characteristics is demonstrated by the high SSIM value, which guarantees that the created image is visually cohesive and correct to the original description.

Similarly, different kinds of text prompts are given to the system, and the results obtained in the image format are depicted in Table 2.

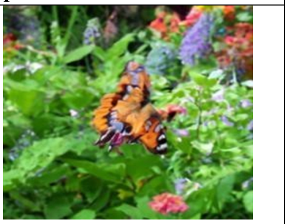
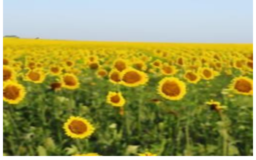
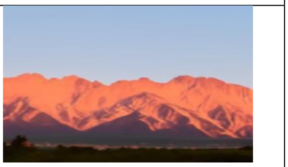
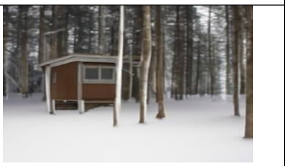
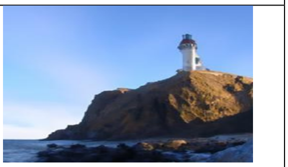
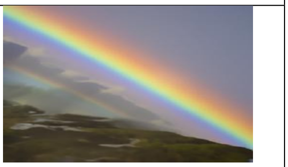
Input: Prompt	Image generated by proposed Model
A colorful garden with butterflies	
A field of blooming sunflowers	
A majestic mountain range at sunset	
A snowy forest with a cabin	
A towering lighthouse on a rocky coast	
A vibrant rainbow over a waterfall	

Table 2. Generated image based on the input prompt.

B. DISCUSSION ON RESULTS

We've employed two distinct image categories within this context: "gt_images" and "gen_images." The former, "gt_images," comprises authentic images sourced from Google, utilizing the same prompt for generating images through the model. Conversely, the latter, "gen_images," denotes images generated by the model. To comprehensively assess the efficacy of our proposed methodology, we employed a set of 50 randomly chosen prompts and procured corresponding images from the Chrome platform. Subsequently, a meticulous evaluation unfolded wherein the generated images were subjected to a stringent comparison against ground truth images, facilitated by the Fréchet Inception Distance (FID) metric. This evaluation's outcomes compellingly advocate for our approach's superiority over prevailing text-to-image generation techniques, as evident from the FID scores.

C. THE PROPOSED MODEL EFFICIENCY:

The proposed model efficiency is calculated by using the following indicators:

1) The "Peak Signal-to-Noise Ratio (PSNR)" and
2) The "Structural Similarity Index (SSIM)".

The first efficiency indicator is the Peak signal-to-noise ratio (PSNR) [27]. This metric is determined by contrasting the highest pixel intensity against the mean squared error (MSE) of the two images in consideration. A superior PSNR value signifies a nearer match between the generated image and its corresponding textual description. The quantitative measure of the ratio between the noise signal power to the original signal's maximum power is termed as the Peak Signal to Noise Ratio (PSNR). If `gt_image` is the real image, and denoted by gr and `gen_image` is the generated image denoted by gg, then the PSNR can be calculated as follows:

The mean square error (MSE) is defined by

$$\text{MSE} = \text{mean}\left((gr - gg)^2\right) \qquad \text{(Eq 1)}$$

If MSE $= 0$, no noise is present in the image, else

$$\text{PSNR} = 20 \cdot \log_{10}\left(\frac{m_i}{\sqrt{\text{MSE}}}\right) \qquad \text{(Eq 2)}$$

where $m_i = 255.0$.

Structural Similarity Index (SSIM) [28]: This evaluates the similarity in structure between two images by examining their luminance, contrast, and structural attributes. An SSIM score nearing 1 denotes a strong likeness of the generated image to its reference or ground truth. article amsmath

The Structural Similarity Index (SSIM) between two images, gr (reference image) and gg (generated image), is given by:

$$\text{SSIM}(gr, gg) = \frac{(2\mu_{gr}\mu_{gg} + C_1)(2\sigma_{gr,gg} + C_2)}{(\mu_{gr}^2 + \mu_{gg}^2 + C_1)(\sigma_{gr}^2 + \sigma_{gg}^2 + C_2)} \qquad \text{(Eq 3)}$$

where:

$$C_1 = (K_1 \cdot m_i)^2, \quad C_2 = (K_2 \cdot m_i)^2$$

μ_{gr} is the mean intensity of the reference image gr.
μ_{gg} is the mean intensity of the generated image gg.
σ_{gr}^2 is the variance of the reference image gr.
σ_{gg}^2 is the variance of the generated image gg.
$\sigma_{gr,gg}$ is the covariance of gr and gg.
K_1 and K_2 are constants, typically $K_1 = 0.01$ and $K_2 = 0.03$.

m_i is the dynamic range of the pixel values (usually 255 for 8-bit images). article amsmath

The Mean SSIM index over a specific window size W is given by:

$$\text{MSSIM} = \frac{1}{M}\sum_{j=1}^{M} \text{SSIM}(gr_j, gg_j) \qquad \text{(Eq 4)}$$

where:
- M is the number of windows in the image. - SSIM(gr_j, gg_j) is the SSIM index calculated for the j-th window of the reference image gr_j and the generated image gg_j.

Table 3. Model efficiency scores

SNo	Parameter	Obtained Average Score
1	PSNR	183.66
2	SSIM	0.9999

Furthermore, the structural integrity of our model was rigorously gauged through applying the Structural Similarity Index (SSIM). Impressively, our model garnered a commendable SSIM score of 0.999, as depicted in Table 3, underscoring its remarkable precision in generating images from arbitrary textual prompts. The quality of the generated images attests to their excellence, with a coherent semantic alignment to the provided input text. Notably, our proposed approach shines in its generative prowess and computational efficiency, setting it apart from GAN-based counterparts.

This efficiency is poised to significantly streamline the image synthesis process, augmenting the viability of our approach for practical applications in real-world scenarios.

The detailed efficiency scores of the proposed model obtained during implementation in Google Colab are represented in Table 3.

D. PROPOSED MODEL LIMITATIONS:

Despite extensive filtering of the pre-training dataset, GLIDE (filtered) remains susceptible to biases that extend beyond those observed in images featuring individuals. In this investigation, we delve into several instances of these biases, illuminating their presence:

In response to requests for generating toys tailored for both genders, GLIDE (filtered) exhibits variations in its outputs. When prompted to generate imagery depicting "a religious place," the model tends to generate images portraying churches predominantly. Notably, the influence of classifier-free guidance amplifies this inclination. Beyond its capacity to generate swastikas and confederate flags, GLIDE (filtered) may exhibit an increased propensity to produce images that bear semblance to hate symbols. Regrettably, the filtration process aimed at hate symbols was predominantly focused on These two emblematic cases are due to the limited availability of pertinent images in our dataset. Alarmingly, this has led to a noticeable decline in the model's proficiency in handling a broader array of symbols.

VI. CONCLUSION AND FUTURE SCOPE

The proposed diffusion model has showcased remarkable potential in text-to-image synthesis, yielding exceptional visual outputs from textual descriptions. Recent strides in this domain have significantly elevated image quality, with the advent of innovative techniques such as two-stage diffusion, diversity regularization, and progressive refinement, all aimed at bolstering the prowess of the diffusion model. However, while progress has been substantial, challenges persist on the horizon. A pivotal hurdle involves the creation of an expansive spectrum of lifelike visuals capable of faithfully representing the provided textual narratives. Additionally, the model's scalability remains a paramount concern, particularly concerning high-resolution images that can lead to unwieldy image sizes and an extensive proliferation of model parameters. It demonstrates outstanding improvement in the quality of the image with an average of 183.66 dB and an SSIM of 99.99%, showing its efficacy in resulting in photorealistic images from textual captions.

This research work in the future holds great promise for text-to-image synthesis using the diffusion model. Future attempts may include developing unique approaches for producing a greater range of realistic and diversified visuals. Addressing the scalability problem requires investigation, with efforts to fulfill high-resolution demands efficiently.

References

[1] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," Advances in neural information processing systems, vol. 34, pp. 8780–8794, 2021.

[2] J. Ho and T. Salimans, "Classifier-free diffusion guidance," arXiv preprint arXiv:2207.12598, 2022.

[3] B. Dayma and P. Cuenca, "Dall. e mini-generate images from any text prompt," Weights & Biases, 2022.

[4] S. Geng, J. Yuan, Y. Tian, Y. Chen, and Y. Zhang, "Hiclip: Contrastive language-image pretraining with hierarchy-aware attention," arXiv preprint arXiv:2303.02995, 2023.

[5] R. T. Hughes, L. Zhu, and T. Bednarz, "Generative adversarial networks–enabled human–artificial intelligence collaborative applications for creative and design industries: A systematic review of current approaches and trends," Frontiers in artificial intelligence, vol. 4, p. 604234, 2021.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[7] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in International conference on machine learning. PMLR, 2016, pp. 1747–1756.

[8] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in International conference on machine learning. PMLR, 2016, pp. 1060–1069.

[9] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5907–5915.

[10] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1316–1324.

[11] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1505–1514.

[12] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18 208–18 218.

[13] D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba, "Gan dissection: Visualizing and understanding generative adversarial networks," arXiv preprint arXiv:1811.10597, 2018.

[14] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in Proceedings of the IEEE/CVF

conference on computer vision and pattern recognition, 2022, pp. 2426–2435.

[15] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," arXiv preprint arXiv:1809.11096, 2018.

[16] W. Dong, S. Xue, X. Duan, and S. Han, "Prompt tuning inversion for text-driven image editing using diffusion models," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7430–7440.

[17] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6007–6017.

[18] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent progress on generative adversarial networks (gans): A survey," IEEE access, vol. 7, pp. 36 322–36 333, 2019.

[19] K. P. Murphy, Machine learning: a probabilistic perspective. MIT press, 2012.

[20] J. Liu and T. H. Lin, "A framework for the synthesis of x-ray security inspection images based on generative adversarial networks," IEEE Access, vol. 11, pp. 63 751–63 760, 2023.

[21] Z. Zhang, L. Han, A. Ghosh, D. N. Metaxas, and J. Ren, "Sine: Single image editing with text-to-image diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6027–6037.

[22] P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 37–45.

[23] P. Agrawal, R. Girshick, and J. Malik, "Analyzing the performance of multilayer neural networks for object recognition," in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13. Springer, 2014, pp. 329–344.

[24] M. Lučić, M. Tschannen, M. Ritter, X. Zhai, O. Bachem, and S. Gelly, "High-fidelity image generation with fewer labels," in International conference on machine learning. PMLR, 2019, pp. 4183–4192.

[25] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," arXiv preprint arXiv:2112.10741, 2021.

[26] R. T. Whitaker and S. M. Pizer, "A multi-scale approach to nonuniform diffusion," CVGIP: image understanding, vol. 57, no. 1, pp. 99–110, 1993.

[27] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in 2010 20th international conference on pattern recognition. IEEE, 2010, pp. 2366–2369.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.

DR.T.M.N.VAMSI is working as Associate Professor in the Department of Computer Science and Engineering at GITAM Deemed to be University, Visakhapatnam, Andhra Pradesh. He received his PhD in Computer Science and Engineering from JNTUH, Hyderabad in the year 2016. He has 24 years of teaching, research, and administrative experience in various technical higher education Institutions. His research interests are in the development of protocols for the Internet of Things, Vehicular networks, and actively doing research in Soft Computing and Bioinformatics. He authored 28 research articles in various reputed international, and national journals and conferences. He is a member of IEEE, CSI, and IEI.

J. N.V.R. SWARUP KUMAR is an assistant professor at GITAM (Deemed to be University) Visakhapatnam, India. He has a PhD from the Department of Information Technology, Annamalai University, Tamilnadu, M.Tech. degree in Information Technology from GITAM University, Visakhapatnam, India, and B.Tech. degree in Computer Science and Engineering from Pondicherry University, India. I bring extensive teaching experience and expertise in various fields. My papers have been published in International Journals and Conferences of IEEE, SPRINGER, and ELSEVIER, covering IoT, Low Power Networks, Cloud Computing, Data Science, and Image Processing Techniques.

DR. I SUNDARA SIVA RAO obtained his Ph.D. in Computer Science and Engineering from GITAM-A Deemed-to-be University, Visakhapatnam, India, 2017. He is having 23 years of teaching experience in various Engineering Colleges. Presently he is working as Professor in the Department of Computer Science and Systems Engineering, at Andhra University. He is having more than 16 publications in various reputed international journals viz., IEEE and Springer. He also Published 5 Patents by IP India Services Government of India. He is a life member of ISTE, IEI, and CSTA. His current research interest includes Machine Learning, IoT, and Artificial intelligence.

MRS. L.PRATIBHA is an Assistant Professor at the Department of CSE, Engineering and Technology Programs of GVP College for Degree and PG Courses(A), Visakhapatnam. Currently, she is pursuing her Ph.D. at GIET University and has 13 years of teaching experience. She published a good number of publications in National and International Journals and conferences. She is a life member of IEI, and her current research interest includes Machine Learning and IoT.