iJRASET
International Journal For Research in
Applied Science and Engineering Technology

Volume: 12 Issue: XI Month of publication: November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65677

Evaluating Text-to-Image Generation Methods: Stable Diffusion vs Generative Adversarial Networks (GANs)

Indumathi. D[1], Tharani. S[2]
[1]Associate Professor, [2]PG Scholar, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamil Nadu, India

Abstract: Text-to-image generation is a fast developing field of artificial intelligence that allows verbal descriptions to be converted into realistic or creative visuals. This study investigates the differences between two cutting-edge approaches for generating images from text, examining their performance, efficiency, and practical applicability across multiple areas. The dominant techniques in this discipline are Generative Adversarial Networks (GANs) and Stable Diffusion models. While GANs have long been the preferred architecture for picture generation tasks, newer diffusion-based models such as Stable Diffusion have emerged as viable alternatives, providing distinct methods to noise reduction and image synthesis. Attention GAN(AttnGAN), a GAN-based approach, uses attention mechanisms to improve the semantic alignment of text descriptions and generated images, resulting in more contextually appropriate graphics. These methodologies are compared, with an emphasis on architectural differences, performance, and applicability to varied applications. GANs use adversarial training, in which two networks (the generator and the discriminator) compete to produce increasingly realistic images. This method is quite effective for producing high-quality photos, but it has drawbacks such as mode collapse and training instability. In contrast, Stable Diffusion models use a probabilistic diffusion process to iteratively reduce noisy images into coherent outputs, resulting in increased processing efficiency and the ability to handle high-resolution images. Experimental evaluation of benchmark datasets reveals each method's strengths and limits in real applications such as digital art, content development, and product design. Stable Diffusion produces more diverse and high-resolution images with fewer computer resources, but GANs generate extremely detailed and realistic visuals. The comparative insights gathered from this research can be used to choose the best technique for a given text-to-image production problem.

Keywords: Stable Diffusion, Scheduler, Generator, Discriminator, AttnGAN

I. INTRODUCTION

In recent years, Stable Diffusion and Generative Adversarial Networks (GANs) have emerged as effective methods for creating images from text descriptions. Stable Diffusion uses a diffusion process in which noise is iteratively introduced to an image and then reversed, resulting in high-quality graphics directed by the input text. Its capacity to perform complicated visual tasks and scale across domains makes it extremely successful at creating realistic visuals. In contrast, GANs are made up of two neural networks—a generator and a discriminator—that compete to produce increasingly realistic images. A special form, Attention GAN (AttnGAN), increases image quality by introducing attention methods that allow the model to focus on different parts of a written description during generation.

A. Stable Diffusion

Stable Diffusion is a cutting-edge AI model that marks a huge step forward in text-to-image generation. It allows users to generate high-quality images from text descriptions in a way that mimics natural image generation. This model stands out for its speed, performance, and low computing resource requirements, making it suitable for a wide range of users.

B. Components of Stable Diffusion

Stable Diffusion is a cutting-edge approach for text-to-image conversion that use a probabilistic diffusion process to convert noisy inputs into high-quality images. The main components of stable diffusion are:

The Text Encoder (CLIP Model) converts input text into a numerical representation that encodes word semantics. It generates embeddings, which serve as the basis for the image production process.

The Image Information Creator (UNet + Scheduler) component employs a diffusion technique to generate image data from noise. This is where Stable Diffusion receives a speed advantage by operating in the latent space.
The Image Decoder (Autoencoder) is the final component that turns processed latent data into pixel-based images.

C. Working of Stable Diffusion

Stable Diffusion generates images by iteratively refining noise into meaningful visuals, guided by text embeddings from a text encoder. The noise scheduler adds controlled noise, and the diffusion model denoises it step by step, aligning the output with the input text. This process ensures high-quality, realistic images that accurately reflect the textual description.

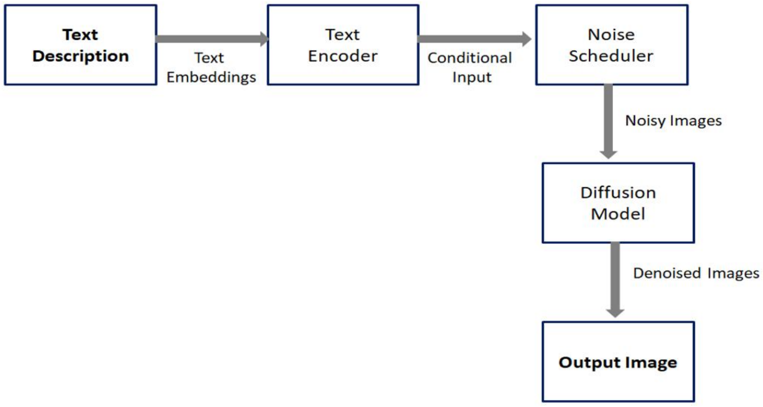


Fig. 1 Working of Stable Diffusion

Figure 1 depicts the process of creating visuals from textual descriptions using a diffusion model. It starts with a text description as input, which is encoded into numerical representations known as text embeddings. These embeddings contain the text's semantic meaning, allowing the model to recognize the essential features and concepts described in the input.
The text embeddings are then passed as conditional input to a noise scheduler. The noise scheduler adds controlled noise to images during the production process, allowing the system to experiment with a range of visual representations. This stage guarantees that the model can generate different and realistic visuals while being consistent with the text.
The scheduler's noisy images are processed using a diffusion model, which iteratively denoises them. At each stage, the model refines the image by reducing noise and increasing details based on the textual input. This technique ensures that the output image is semantically coherent with the input text and produces high-quality visual results.
Finally, the denoising procedure creates a realistic and detailed image that matches the written description. This pipeline highlights the ability of diffusion models in producing visuals that closely match complex verbal descriptions, making them useful for a variety of applications.

D. Applications of Stable Diffusion

Stable Diffusion has transformed the area of digital art and design by enabling artists to create spectacular pictures with basic text cues. This has democratized the art-making process, allowing anyone with little to no drawing skills to produce professional-quality work. AI now allows designers to swiftly develop graphics for branding, marketing, and conceptual art, saving time and resources. The model may even be fine-tuned to produce various styles, ranging from abstract art to realistic portraits, making it an adaptable tool in the creative sectors. In the content creation sector, Stable Diffusion can be used to create unique visuals for blogs, social media, and advertising campaigns. Instead of using generic images, artists can create custom visuals that are consistent with their brand or message. This tool has been very effective for improving visual storytelling in media, enabling for the quick creation of bespoke graphics to supplement written content. For example, news sites or bloggers can utilize it to create relevant visuals for stories, increasing reader engagement. Stable diffusion has the potential to improve medical imaging by allowing for the viewing of anatomical structures or anomalies based on textual descriptions. While the system is now focused on artistic or conceptual images, it might be modified to generate illustrative medical visuals based on doctors' descriptions, adding another layer of aid in diagnosis or education. For example, it might be used in medical education to generate detailed drawings of organs or procedures based on text input of specific illnesses or surgical techniques, making it an effective tool for training and simulation.

E. Generative Adversarial Networks

Generative Adversarial Networks are a type of machine learning model that is used to produce new data instances that are similar to the training data. GANs are made up of two neural networks, a generator and a discriminator, which are trained simultaneously via an adversarial process. The generator generates data that resembles genuine data, while the discriminator determines whether the data is real or produced. The two networks compete, with the generator attempting to improve its ability to produce realistic data and the discriminator aiming to improve its capacity to discriminate between actual and fraudulent data. GANs have grown in popularity because to their capacity to generate realistic images, movies, and even music from random inputs, making them especially helpful for content creation and image synthesis.

F. Types of GAN

The Vanilla GAN is a basic type of Generative Adversarial Network. It comprises of a simple generator and a discriminator, both of which use completely connected layers. The generator generates synthetic data from random noise, and the discriminator determines if the data is real or phony. The goal of training a Vanilla GAN is for the generator to produce data that is indistinguishable from actual data, causing the discriminator to misclassify phony data as real.

Conditional GANs build on the basic GAN paradigm by conditioning the generator and discriminator on extra information, such as class labels or text descriptions. This helps the model to produce more tailored results. For example, rather than generating random images, a cGAN might be trained to generate images of specific things (e.g., dogs, cats, or birds) by feeding class labels into both networks.

DCGANs use convolutional layers in both the generator and the discriminator, which is especially useful for image generation. The convolutional layers enable the model to grasp spatial relationships in the data, resulting in higher image quality than the fully connected networks employed in Vanilla GANs.

CycleGANs are used for image-to-image translation tasks, which involve transforming images from one domain to another without the assistance of matched samples. For example, CycleGANs can be trained to turn photographs of horses into photos of zebras, and vice versa. This model is unique in that it uses a cycle consistency loss to ensure that the transformation is reversible, allowing the translated images to be converted back to their original form.

AttnGAN is a sort of GAN that uses attention techniques to focus on specific parts of the text description during image production. Unlike typical GANs, which may fail to capture all of the nuances from complicated textual descriptions, AttnGAN enables the model to create images with finer detail by focusing on certain words or phrases at various stages of the generation process.

G. Components of GAN

The key components of Generative Adversarial Networks (GANs), emphasizing the interaction between the generator and the discriminator. The generator creates synthetic data samples that resemble real data, while the discriminator examines these samples to distinguish between genuine and produced data.

The generator is in charge of creating phony data that looks like actual data. It takes random noise or latent variables as input and converts them into data that may be included in the training dataset. Its purpose is to "fool" the discriminator into believing that the created data is legitimate. The generator improves with time as it learns to deliver more realistic results.

The discriminator is a classifier that distinguishes between actual data from the training set and fabricated data generated by the generator. It returns a probability score indicating whether the input data is authentic or phony. As the discriminator improves at discriminating between genuine and created data, it assists the generator by finding weaknesses in the phony data.

GANs use an adversarial training setup, pitting the generator and discriminator against one another. The generator's goal is to create data that will trick the discriminator, whereas the discriminator's goal is to accurately discern between actual and produced data. This competitive process helps both networks progress until they reach a point where the discriminator can no longer tell the difference between actual and created data and the generator produces extremely realistic data.

H. Working of GAN

The operational principles of Generative Adversarial Networks (GANs) describe how the generator and discriminator interact in an adversarial setting. The generator generates synthetic data samples designed to closely resemble actual data, while the discriminator compares these samples to real data to establish their authenticity. The training procedure includes a continuous feedback loop in which the generator improves its outputs depending on the discriminator's evaluations, resulting in higher quality generated data over time.

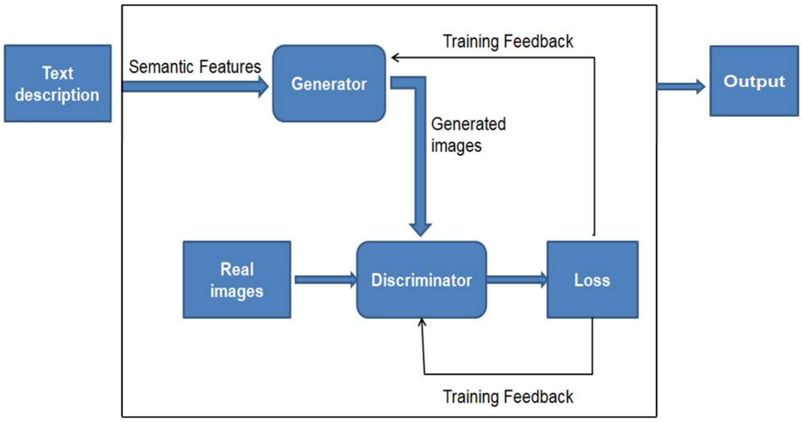


Fig. 2 Working of GAN

1) *Generator Network:* The generator uses random noise as input and seeks to produce data that is similar to the original dataset. For example, in an image generating task, the generator creates images using noise vectors.
2) *Discriminator Network:* The discriminator receives both true data from the original dataset and false data from the generator. Its purpose is to classify data as true or fraudulent by generating a likelihood score.
3) *Adversarial Process:* The generator and discriminator are taught simultaneously, but with opposing purposes. The generator is trained to "fool" the discriminator, but the discriminator is trained to correctly discern between actual and bogus data.
4) *Loss Function Optimization:* The generator seeks to reduce loss (fooling the discriminator), whereas the discriminator seeks to increase performance (properly distinguishing genuine and fake data). Over time, the two networks compete, and their performance improves with iterations.
5) *Equilibrium:* Training continues until the discriminator is unable to distinguish between actual and bogus data. At this stage, the GAN has found equilibrium and the generated data is quite lifelike.

I. Applications of GAN

GANs can create new images using text or low-resolution inputs. They can also modify photographs by adding items or adjusting their properties, such as increasing image resolution or transforming them from black and white to color. GANs create synthetic data to supplement existing datasets, particularly in machine learning. This is especially beneficial in situations such as fraud detection, where creating fake but realistic data enhances model resilience. GANs are used to complete missing data or fill gaps in images. GANs, for example, may predict how different areas of an image would look based on visible information, such as when constructing subterranean maps for energy applications. GANs turn 2D photos into 3D models, which is especially beneficial in industries such as healthcare, where 3D organ models are required for surgical planning purposes. GANs enable to generate high-resolution images from scans and visualize complicated structures, which improves surgical precision and medical diagnostics.

Generative Adversarial Networks (GANs) have a wide range of applications, including image synthesis, video production, and data augmentation. GANs are used in image synthesis to generate high-resolution images, which allows for applications in art generation, fashion design, and virtual reality environments. They are also used to create realistic animations and deepfakes in the video production sector. Furthermore, GANs play an important role in data augmentation for training machine learning models, especially in circumstances with insufficient labeled data, by producing synthetic examples that increase dataset diversity.

J. Objectives

The principal objectives are to investigate and evaluate Generative Adversarial Networks (GANs) and Stable Diffusion.

1) *Evaluate Performance:* Analyze and compare the accuracy, image quality, and consistency of generated images for Stable Diffusion and GAN-based models using quantitative and qualitative measures.
2) *Assess Semantic Alignment:* Examine how well each technique fits the generated images with the related text descriptions, emphasizing strengths and limitations in collecting textual details.
3) *Explore Model Efficiency:* Examine Stable Diffusion and GANs' computational efficiency, resource usage, and training stability, with a focus on training duration, memory consumption, and scalability.

4) *Identify Use Cases:* Identify practical applications where Stable Diffusion or GANs outperform, such as fine-grained picture production, creative image creation, or photo-realistic synthesis.
5) *Suggest Optimizations:* Propose potential optimizations for both Stable Diffusion and GANs based on observed limits, providing insights for future text-to-image synthesis models.

II. LITERATURE SURVEY

A novel method for creating context-aware images with generative artificial intelligence is given. The authors concentrate on using prompt editing to improve contextual coherence, addressing the significant difficulty of maintaining meaning across numerous sentences in image production challenges. The results show that the suggested model improves picture similarity by 30% and ROUGE-recall by 130% when compared to existing models, proving its ability to maintain contextual relevance. Despite these advances, the model struggles in instances with complicated contexts in which numerous sentences must be processed concurrently[1].

The findings highlight the necessity of producing not only realistic images, but also those that accurately reflect the intended context, which is critical for applications that require high levels of semantic accuracy in created material. This novel approach provides new opportunities for improving text-to-image synthesis, emphasizing the importance of striking a balance between computational efficiency and contextual integrity[2].

The study contributes significantly to the field of generative models by giving techniques for improving the quality of generated images in comparison to their descriptive texts. The authors provide a Diverse Conditional GAN (DG-VRT) framework for improving visual representation from text descriptions in visual recognition and semantic segmentation challenges. The system uses Deep Convolutional GANs (DCGANs) to produce high-quality, diversified synthetic images that cater to a variety of recognition scenarios[3].

The suggested paradigm significantly improves visual identification performance, especially in complicated situations with intricate object interactions. However, problems develop when the model is confronted with several objects and their interactions, reducing the quality of the generated images. The importance of diversity in picture production, arguing that diverse visual outputs are required to improve machine learning tasks involving visual recognition. This method emphasizes the importance of systems that can effectively analyze and generate images from complicated textual inputs, while also providing insights on how to improve the resilience of visual recognition frameworks. Overall, the work advances picture synthesis techniques by highlighting the interplay of text and visual information [4].

Optimal performance requires synchronization between the generator and discriminator, as inconsistencies might cause mode collapse or reduction in sample quality. This essential work provides the framework for future advances in adversarial learning, providing important insights into the mechanisms that enable effective data production. The ideas presented in this study continue to inspire current research in a variety of fields, emphasizing GANs' adaptability and potential in synthetic data production [5].

However, the model is constrained by its reliance on synthetic data; future research should focus on improving real-world image synthesis and caption generation for a broader range of settings. A novel architecture is proposed, which employs context-aware cross-attention alignment and adversarial learning to generate different visuals led by exemplars. The model successfully creates realistic and semantically consistent images by adding style components from example photos. This end-to-end approach allows for the creation of different photos with guided style aspects, increasing the realism of created content. However, the complexity of training both the text and image encoders poses challenges, and the model may become cumbersome as the diversity of exemplars increases[6].

The authors provide a new GAN architecture aimed at reliably creating distinct items based on textual descriptions. The research introduces a new evaluation metric, Semantic Object Accuracy (SOA), that is more consistent with human evaluations of object accuracy in generated images [7]. The model improves detail and accuracy in object production, particularly in compared to earlier GAN architectures. While the model considerably increases item accuracy, it still struggles to generate complex objects and environments, especially when dealing with uncommon or unknown objects[8].

The model exceeds many cutting-edge GANs in creating photorealistic images across multiple categories. However, balancing the many constraints inside the model adds complexity, and the model's performance varies depending on the dataset employed.A innovative architecture that uses Dual Injection Blocks and Efficient Conditional Channel Attention (ECCA) modules to improve text-to-image diversity and efficiency. DE-GAN successfully generates diverse and realistic graphics with minimum storage overhead, making it ideal for resource-constrained applications[9].

A deep GAN architecture is proposed that generates believable images from comprehensive text descriptions. The model uses a manifold interpolation regularizer to increase synthesis quality. The results show that the model can produce very convincing images on datasets like CUB and MS-COCO. While the approach successfully separates design and content, expanding the model to higher resolutions and a wider range of text inputs remains a challenge. The model shows generalizability across different backdrops and object types[10].

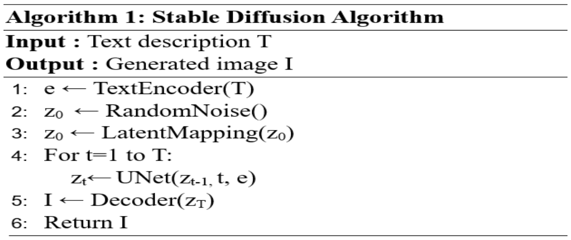
A self-supervised technique (SS-TiGAN) is developed for improving text-to-image synthesis. The approach employs a bi-level architecture with two discriminators, as well as rotation versions, to address GAN stability difficulties. Even in low-data regimes, the model produces high-quality images, as demonstrated by its robustness on datasets such as Oxford-102 and CUB [11]. The model's success is restricted by its output resolution of 128 × 128 pixels, requiring additional refinement for higher-resolution production. The model generates fine-grained images with fewer parameters by utilizing a cross-modal correlation technique and a hinge loss function. The results indicate considerable gains in creating high-resolution photographs with detailed characteristics. However, future research could focus on combining more complex attention mechanisms to improve visual quality and scale[12].

The new dataset for text-to-face tasks, however, is constrained by its size and focus on face photos, leaving potential for future research to improve the detail and diversity of created faces[13]. Conditional GANs (C-GAN), attention processes, and contrastive learning are used to improve image quality and diversity. The model achieves cutting-edge performance, outperforming previous models on the COCO-Stuff dataset. However, the model may need significant computer resources for training, and future research should look into ways to improve its efficiency without sacrificing performance[14]. AttnGAN is presented as a unique framework that uses attention techniques to focus on specific words during image production. This multi-stage refinement model beats earlier techniques, especially for producing fine-grained features. However, the intricacy of attention mechanisms raises processing requirements, therefore future research could focus on increasing training efficiency[15].

III. IMPLEMENTATION

The implementation for Stable Diffusion and Generative Adversarial Networks (GANs) generate images from text descriptions using the Stable Diffusion model and GAN model, which includes text encoding, latent space manipulation, and iterative denoising.

A. *Pseudocode for Stable Diffusion Model*

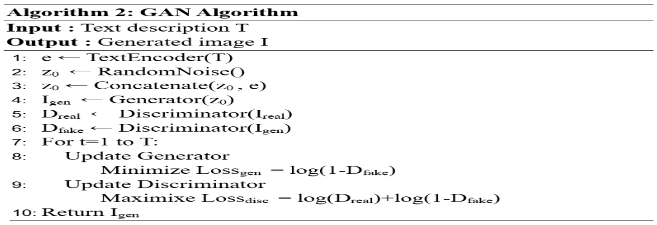


```
Algorithm 1: Stable Diffusion Algorithm
Input : Text description T
Output : Generated image I
e ← TextEncoder(T)
$z_0$ ← RandomNoise()
$z_0$ ← LatentMapping($z_0$)
For t=1 to T:
    $z_t$← UNet($z_{t-1,}$ t, e)
I ← Decoder($z_T$)
Return I
```

Fig. 3 Pseudocode for Stable Diffusion

The pseudocode in Fig 3 for the Stable Diffusion algorithm creates images from text descriptions using a pre-trained text encoder, latent mapping, denoising, and image decoding models. Initially, the text encoder converts the text description into a feature vector. A random noise vector is created, transferred to the latent space, and repeatedly tuned using the UNet model across several diffusion timesteps. The UNet gradually denoises the latent representation, depending on both the initial noise and the encoded text properties. Finally, the corrected latent representation is decoded into a produced image, which serves as the output.

B. *Pseudocode for GAN Model*



```
Algorithm 2: GAN Algorithm
Input : Text description T
Output : Generated image I
e ← TextEncoder(T)
$z_0$ ← RandomNoise()
$z_0$ ← Concatenate($z_0$ , e)
$I_{gen}$ ← Generator($z_0$)
$D_{real}$ ← Discriminator($I_{real}$)
$D_{fake}$ ← Discriminator($I_{gen}$)
For t=1 to T:
    Update Generator
        Minimize $Loss_{gen}$ = log(1-$D_{fake}$)
    Update Discriminator
        Maximixe $Loss_{disc}$ = log($D_{real}$)+log(1-$D_{fake}$)
Return $I_{gen}$
```

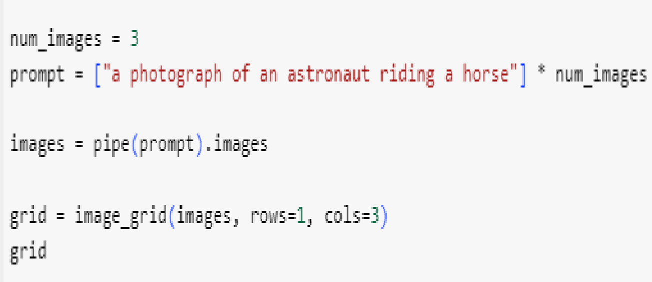
Fig. 4 Pseudocode for GAN

Figure 4 shows the pseudocode for the GAN algorithm for text-to-image generation, which begins by encoding the input text description with a text encoder and then combining it with random noise to form a latent vector. The generator use this latent vector to generate a synthetic image. The discriminator assesses both real and produced images to determine whether they are real or fraudulent. The generator is updated by minimizing the loss based on the discriminator's ability to identify the generated image as fake, and the discriminator is updated by increasing its ability to discriminate between real and fake images. This adversarial process continues until the generator creates realistic visuals that are indistinguishable from the genuine thing.

IV. RESULTS AND DISCUSSION

The results of using the Stable Diffusion and GAN models for text-to-image generation displays the generated images from both models and evaluating their quality, fidelity, and diversity using standard evaluation measures like the Inception Score (IS) and Fréchet Inception Distance (FID).

A. Sample Output Using Stable Diffusion



```
num_images = 3
prompt = ["a photograph of an astronaut riding a horse"] * num_images

images = pipe(prompt).images

grid = image_grid(images, rows=1, cols=3)
grid
```

Fig. 5 Prompt in Stable Diffusion

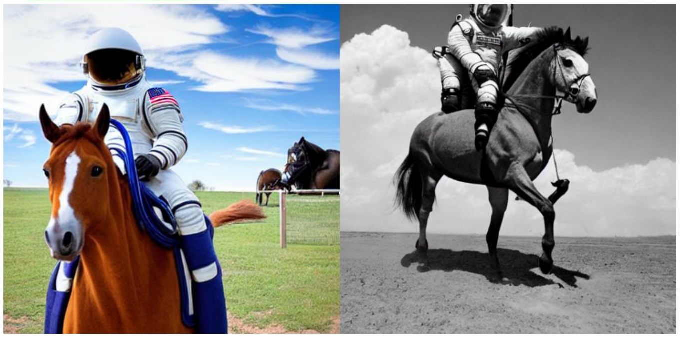
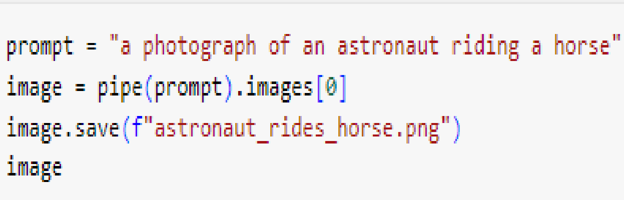


Fig. 6 Output using Stable Diffusion

Figures 5 and 6 show the output of Stable Diffusion for a particular prompt, which is a created image that visually represents the description supplied in the prompt. The model employs advanced diffusion algorithms to convert the text input into a coherent, contextually relevant image.

B. Sample Output Using GAN



```
prompt = "a photograph of an astronaut riding a horse"
image = pipe(prompt).images[0]
image.save(f"astronaut_rides_horse.png")
image
```

Fig. 7 Prompt in GAN

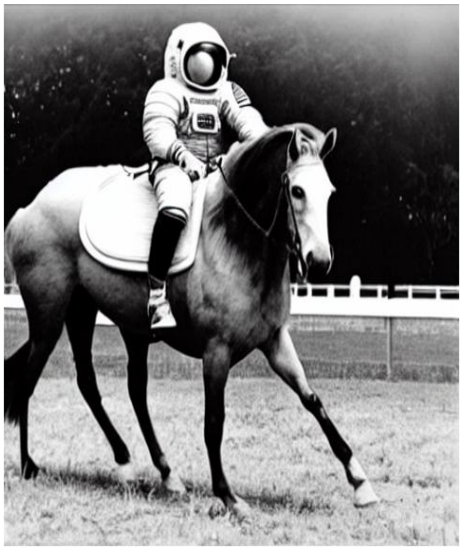


Fig. 8 Output using GAN

Figures 7 and 8 show the output of a Generative Adversarial Network (GAN) for a given prompt, which is a produced image that strives to resemble real images based on training data.

C. Performance Evaluation of Generated Images

1) Inception Score (IS)

The Inception Score measures the quality and diversity of created images. It uses a pre-trained Inception model to classify the generated images. The score is derived by comparing the conditional probability of class labels given the generated photos to the marginal probability of class labels over all images. A higher Inception Score implies that the generated images are not only visually appealing, but also diversified in content, since they should belong to different categories and have high-quality attributes.

$$I = exp(\mathbb{E}_x D_{KL} (p(y|x)||p(y))) \qquad (1)$$

The difference between the marginal distribution p(y) and the conditional distribution p(y|x) is calculated by IS using the Kull back-Leibler (KL) divergence. The generated image x, represented by the label y, is predicted using a pre-trained Inception v3 network.

2) Fréchet Inception Distance (FID):

FID assesses the resemblance between the distribution of generated and real images in feature space. It computes the difference between the mean and covariance of feature representations (obtained from a pre-trained Inception model) for actual and produced images. Lower FID ratings indicate that the generated images are closer to the original images, implying higher fidelity and quality. FID is regarded more robust than other measures such as IS since it considers feature distribution rather than merely probabilities.

$$F(r,g) = ||\mu_r - \mu_g||^2 + trace\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}\right) \qquad (2)$$

where r and g denote the image's real and produced characteristics. The covariance and mean of actual and generated features are denoted as r, g, r, and g, respectively.

3) Dataset

The Flickr 8K dataset is commonly used for applications like picture captioning, image-to-text production, and text-to-image conversion. It comprises of 8,000 photos, each with five informative descriptions, for a total of 40,000 captions. The photographs come from the Flickr platform and depict a variety of ordinary scenarios, such as people, animals, objects, and landscapes.

4) Data Samples

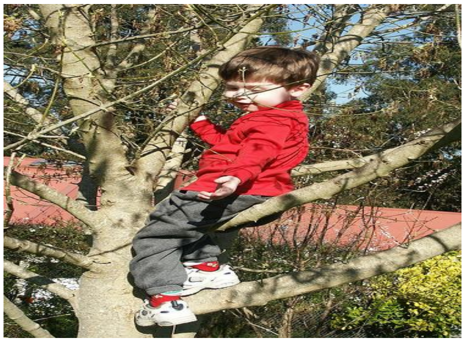
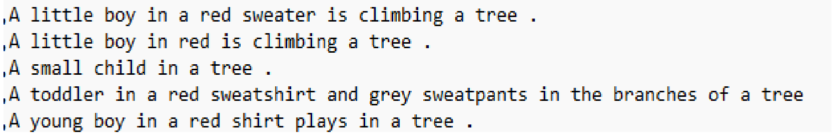


Fig. 9 Image Sample



```
,A little boy in a red sweater is climbing a tree .
,A little boy in red is climbing a tree .
,A small child in a tree .
,A toddler in a red sweatshirt and grey sweatpants in the branches of a tree
,A young boy in a red shirt plays in a tree .
```

Fig. 10 Caption Sample

In Figures 9 and 10, an image is shown with five descriptive subtitles. Each caption stresses a distinct facet of the image, resulting in a diverse textual depiction of its substance. The use of images and captions displays how written information can describe visual aspects in a variety of ways.

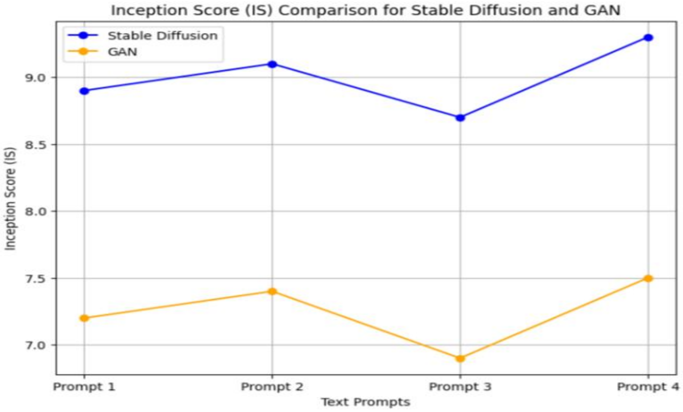


Fig. 11 IS Score Comparison

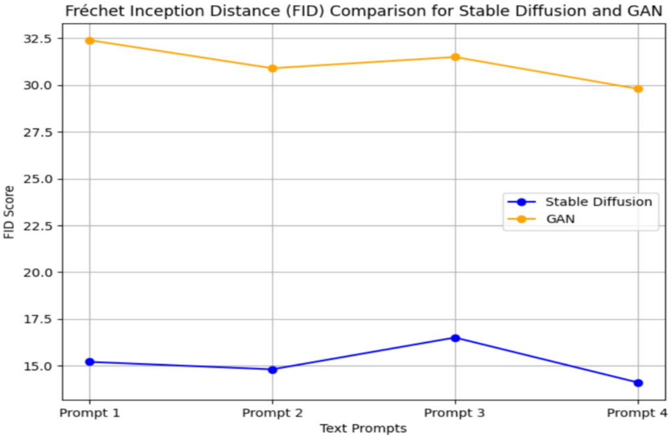


Fig. 12 FID Score Comparison

Figure 9 depicts an IS score graph comparing the performance of Stable Diffusion and GAN across four different text prompts. Stable Diffusion consistently achieves better IS ratings, indicating that it produces more realistic and diversified images. This shows that Stable Diffusion is better at producing distinct, high-quality images from text descriptions, but GAN performs less well in terms of realism and diversity. IS is important since it considers both the clarity of individual images and the number of distinct images that a model may generate.

The FID score graph in Figure 10 depicts how the Stable Diffusion and GAN models perform in terms of producing images that are similar to real-world images. Stable Diffusion consistently has lower FID scores, indicating that its images are more closely distributed to genuine images and thus of higher quality. GAN, on the other hand, provides higher FID scores, implying that the generated images are less realistic than Stable Diffusion. The value of FID lies in determining how well the generated picture set matches the genuine image set, with lower scores indicating more accurate and realistic results.

The IS (Inception Score) and FID (Fréchet Inception Distance) are two often used metrics for assessing the quality of generative models, notably for text-to-image generation. IS assesses the diversity and semantic relevance of generated images to the input text, maintaining consistency with the written description. FID evaluates the realism of generated images by comparing their distribution to genuine images, resulting in a quantifiable measure of fidelity. These measures provide complimentary insights by balancing diversity, alignment, and realism, making them appropriate for assessing text-to-image models.

The IS score assesses a model by calculating the entropy of class predictions for generated images made with a pre-trained classifier, rewarding strong confidence and diversity in the generated content. FID computes the Wasserstein distance between the feature distributions of generated and real images in a pre-trained network's latent space, which measures how similar the two distributions are. Lower FID values imply greater resemblance to real images, whereas higher IS scores suggest greater diversity and alignment, offering a complete assessment of the model's performance.

TABLE I
IS SCORES AND FID SCORES OF STABLE DIFFUSION AND GAN MODELS ON FLIKR 8K DATASET

Model	**IS**	**FID**
Stable Diffusion	19.5	14.6
GAN	11.3	33.8

In table 1, Stable Diffusion has a better IS Score because of its capacity to generate diverse and semantically accurate images that correspond to text descriptions. GAN models, such as AttnGAN, often have a lower IS Score but a higher FID Score due to problems such as mode collapse and less stable training as compared to diffusion models.

V. CONCLUSION AND FUTURE WORK

This comparison analysis compares two popular text-to-image generating techniques: Stable Diffusion and GANs, specifically AttnGAN. Stable Diffusion, with its diffusion-based technique, excels at producing high-quality, detailed images by iteratively reducing noise. Its scalability and stability make it appropriate for a wide range of practical applications.

Stable Diffusion can be used in a variety of practical applications, including painting and graphic design, where artists can create unique visuals based on textual cues, allowing for quick concept creation. In marketing and advertising, it permits the production of personalized images for campaigns based on certain themes or messages. Furthermore, Stable Diffusion can improve product visualization by producing realistic representations of products from descriptions, allowing designers to showcase concepts without the use of physical prototypes. In education, it can be used to provide interesting visual content that supplements instructional resources.

Generative Adversarial Networks (GANs) have a wide range of applications, including image enhancement and restoration, which can improve the quality of low-resolution photos and eliminate noise from photographs. In the fashion business, GANs are used to produce new clothing patterns and simulate how items would look on models, expediting the design process. They are also employed in medical imaging to generate synthetic images that aid in training diagnostic algorithms, resulting in improved performance with less real-world data. GANs can also be used in video game creation to create realistic textures and settings, as well as in film production for visual effects and animation.

Future study could focus on merging the strengths of both methods, such as AttnGAN's attention mechanisms and Stable Diffusion's diffusion process. This hybrid technique could result in improved visual fidelity while maintaining semantic accuracy. Furthermore, increasing the computational efficiency of these models and investigating domain-specific modifications, such as medical imaging or art generation, could broaden their utility. Experimenting with larger datasets and fine-tuning models for more diverse linguistic inputs might also help to enhance robustness and generalizability across different text descriptions.

REFERENCES

[1] H. Kim, J. -H. Choi and J. -Y. Choi, "A Novel Scheme for Generating Context-Aware Images Using Generative Artificial Intelligence," in IEEE Access, vol. 12, pp. 31576-31588, 2024, doi: 10.1109/ACCESS.2024.3368871.

[2] T. Hu, C. Long and C. Xiao, "A Novel Visual Representation on Text Using Diverse Conditional GAN for Visual Recognition," in IEEE Transactions on Image Processing, vol. 30, pp. 3499-3512, 2021, doi: 10.1109/TIP.2021.3061927

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza et al., "Generative Adversarial Networks," in arxiv.org, 2014, https://arxiv.org/abs/1406.2661

[4] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga and M. Bennamoun, "Text to Image Synthesis for Improved Image Captioning," in IEEE Access, vol. 9, pp. 64918-64928, 2021, doi: 10.1109/ACCESS.2021.3075579.

[5] A. Li, et al.,"Specific Diverse Text-to-Image Synthesis via Exemplar Guidance" in IEEE MultiMedia, vol. , no. 01, pp. 1-9, 5555, 2024, doi: 10.1109/MMUL.2024.3421243.

[6] T. Hinz, S. Heinrich and S. Wermter, "Semantic Object Accuracy for Generative Text-to-Image Synthesis" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 44, no. 03, pp. 1552-1565, 2022, doi: 10.1109/TPAMI.2020.3021209

[7] Hu, T., Long, C. & Xiao, C. ,"CRD-CGAN: category-consistent and relativistic constraints for diverse text-to-image generation," in Frontiers of Computer Science, 2024, https://doi.org/10.1007/s11704-022-2385-x

[8] Jiang, B., Zeng, W., Yang, C. et al.,"DE-GAN: Text-to-image synthesis with dual and efficient fusion model,"in Multimedia Tools and Applications, 23839–23852, 2024, https://doi.org/10.1007/s11042-023-16377-8

[9] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, Honglak Lee," Generative Adversarial Text to Image Synthesis," in arxiv.org, 2016, https://doi.org/10.48550/arXiv.1605.05396

[10] Y. Yang, L. Wang, D. Xie, C. Deng and D. Tao, "Multi-Sentence Auxiliary Adversarial Networks for Fine-Grained Text-to-Image Synthesis," in IEEE Transactions on Image Processing, vol. 30, pp. 2798-2809, 2021, doi: 10.1109/TIP.2021.3055062.

[11] Y. X. Tan, C. P. Lee, M. Neo, K. M. Lim and J. Y. Lim, "Enhanced Text-to-Image Synthesis With Self-Supervision," in IEEE Access, vol. 11, pp. 39508-39519, 2023, doi: 10.1109/ACCESS.2023.3268869.

[12] R. Li, N. Wang, F. Feng, G. Zhang and X. Wang, "Exploring Global and Local Linguistic Representations for Text-to-Image Synthesis," in IEEE Transactions on Multimedia, vol. 22, no. 12, pp. 3075-3087, Dec. 2020, doi: 10.1109/TMM.2020.2972856.

[13] M. Z. Khan et al., "A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks," in IEEE Access, vol. 9, pp. 1250-1260, 2021, doi: 10.1109/ACCESS.2020.3015656.

[14] M. A. Habib et al., "GACnet-Text-to-Image Synthesis With Generative Models Using Attention Mechanisms With Contrastive Learning," in IEEE Access, vol. 12, pp. 9572-9585, 2024, doi: 10.1109/ACCESS.2023.3342866.

[15] T. Xu, et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018 pp. 1316-1324,doi: 10.1109/CVPR.2018.00143