Advancements in Text-to-Image Generation: A Comparative Study of Model Architectures, Datasets, and Performance Metrics

Tejas Goyal[1,*,†], Kaveesh Khattar[1,†], Kubtoor Patel Dhruv[1,†], Aditya Hombal[1,†] and Mamatha Hosalli Ramappa[1,†]

[1]*Computer Science and Engineering, PES University, 100 Feet Ring Road BSK III Stage, Bangalore, PO-560085, Karnataka, India*

Abstract

Text-to-image creation is a fast expanding topic that has received a lot of attention in the last few years. This study provides a thorough comparative examination of cutting-edge text-to-image generation models, with the goal of providing an overview of their improvements and capabilities. The investigation focuses on the various model architectures, datasets utilised for training and assessment, and performance measures used to assess picture creation quality. Researchers and practitioners may get significant insights into the strengths and shortcomings of different techniques by comparing and contrasting these models, allowing informed decision-making for picking the best text-to image generating model for certain applications.

Keywords

Image Models, Image Processing, Text-to-Image, Generative AI, GAN

1. Introduction

Text-to-image and image-to-text creation [1, 2] is becoming very popular because of its vast use. The goal of this comparison analysis is to identify the advantages and disadvantages of various text-to-image creation techniques [3]. We may learn about the underlying mechanisms that contribute to their picture synthesis skills by investigating their architectural designs. Cogview (ELBO), discrete variational auto-encoders (dVAE), multi-stage AttnGAN, generative adversarial networks (GANs), LSTM+GAN, CycleGAN+BERT, DF-GAN, MirrorGAN, VQ-SEG (a modified VQ-VAE), StackGAN+fine-tuned BERT text encoding models, and DALL-E-2 are among the models investigated. We look at the datasets used by these models for training and assessment in addition to architectural comparisons. This includes well-known benchmarks like as COCO and CUB, as well as bespoke datasets created expressly for text-to-image creation [4]. The diversity and quantity of these datasets, as well as any pre-processing techniques used, have a significant impact on model performance. Various performance indicators have been used in the field to analyse the quality of produced photographs. Our study incorporates

ACI'23: Workshop on Advances in Computational Intelligence at ICAIDS 2023, December 29-30, 2023, Hyderabad, India

*Corresponding author.

†These authors contributed equally.

jaz.goyal@gmail.com (T. Goyal); kaveeshkhattar@gmail.com (K. Khattar); kpdhruvin@gmail.com (. P. Dhruv); hombaladitya30@gmail.com (A. Hombal); mamathahr@gmail.com (M. H. Ramappa)

human assessments, user research, and additional qualitative evaluations that the analyzed models employed, along with perceptual similarity metrics like Frechet Inception Distance and Inception Score. This allows for a thorough assessment of each model's visual accuracy and realism. We hope that this comparative analysis will give scholars and practitioners a full grasp of the various text-to-image creation techniques. We provide vital insights for making educated decisions in picking the most appropriate model for certain applications by emphasising the strengths and drawbacks of each model based on architectural choices, dataset utilisation, and performance indicators. We give a deep study of the model designs, datasets, and performance indicators in the next sections of this work, as well as a comprehensive comparison analysis. We end by summarising the important findings and outlining potential future research avenues in text-to-image creation.

2. Text to Image Models

Text-to-image creation is a difficult problem that seeks to translate verbal descriptions into aesthetically realistic and semantically consistent images automatically. This task is critical in a variety of applications, including computer vision [5], multimedia content generation, and virtual reality. The objective is to bridge the gap between natural language and visual representations, making it possible for robots to interpret and produce visual content. Several models have been created to do this, including older approaches such as cogview and dVAE, as well as cutting-edge techniques such as different GAN models and BERT. These models use largescale picture datasets like MSCOCO, CUB, and Oxford 102 to understand the relationship between written descriptions and visual representations. These models help to improve human machine interactions and facilitate creative content development by creating high-quality visuals that correspond to the provided text. This review lays the groundwork for a more in-depth examination and comparison of various models in the next sections of this study. Brief Introduction to Models:

1. **Cogview:** Cogview is a state-of-the-art text-to-image generating model that combines cognitive theories and deep learning methods. To generate aesthetically consistent pictures from verbal descriptions, it employs attention processes and generative adversarial networks (GANs) [6].
2. **dVAE:** dVAE (disentangled Variational Autoencoder) is a novel model that uses variational autoencoders to disentangle several aspects of variation in pictures. This gives the model more control over the generation process, allowing it to generate various and relevant visuals based on text input.
3. **Multi-Stage AttnGAN:** Multi-Stage AttnGAN is a multistage attention-based GAN model that refines the produced pictures gradually. It uses a hierarchical structure to collect both global and local picture data, resulting in highquality images that match the specified text descriptions.
4. **LSTM+GAN:** To produce visuals from text, LSTM+GAN combines long short-term memory (LSTM) networks with GANs. The LSTM component makes it easier to model sequential information in text, while the GAN component guarantees that the produced pictures are both visually appealing and semantically appropriate.

5. **CycleGAN+BERT:** CycleGAN+BERT is a sophisticated image-to-image translation model that combines CycleGAN with BERT, a pre-trained language model. This paradigm facilitates cross-modal translation between textual descriptions and visual representations by using the bidirectional link between text and images.
6. **GAN:** GAN (Generative Adversarial Network) is a fundamental paradigm for text-to-image generation. It consists of a discriminator network and a generator network that engage in competition during training. Eventually, the generator produces realistic images from text by learning to create images that deceive the discriminator.
7. **DF-GAN:** Deep Fusion Generative Adversarial Network (DF-GAN) is a GAN variation that uses deep fusion methods to collect fine-grained features during picture production. It intends to generate high-resolution pictures with increased visual quality and semantic coherence.
8. **MirrorGAN:** MirrorGAN makes use of an innovative mirrored approach to improve the alignment of text and picture elements. It employs a two-stage generating process, with the first focusing on global coherence and the second on local details, resulting in aesthetically appealing visuals.
9. **VQSEG:** VQ-SEG (Vector Quantized Variational Autoencoder with Semantic Expansion and Geometric Constraints) is a model that combines vector quantization, variational autoencoders, and semantic expansion techniques. It guarantees that the produced pictures have both semantic consistency and visual quality, making them true to the written descriptions supplied.
10. **StackGAN:**StackGAN is a two-step stacked generative adversarial network that creates pictures. The first step creates low-resolution pictures based on text descriptions, which are then refined to produce high-resolution images with increased details and realism.
11. **Dalle2:** Dalle2 is a DALL-E model variation that combines transformers with VQ-VAE (Vector Quantized Variational Autoencoder). It excels at producing very different and imaginative pictures based on text input, providing a wide range of text-to-image conversion options. These models under consideration provide a broad variety of strategies and approaches for text-to-image creation, each with its own set of strengths and qualities.

3. Datasets

Here are summaries of the mentioned datasets:

1. **YFCC100M (Yahoo Flickr 100 Million Creative Commons):** is a huge dataset that contains 100 million Flickr photographs and videos. It is freely distributed under the Creative Commons licence, making it an excellent resource for computer vision and multimedia research. The dataset has been utilised for picture categorization, object identification, and deep learning applications, allowing for breakthroughs in visual perception and analysis.
2. **Microsoft Common Objects in Context (MS-COCO):** The MS-COCO benchmark dataset is commonly used for object identification, segmentation, and captioning tasks. It includes almost 200,000 photos that have precise annotations such as object bounding

boxes, segmentation masks, and image descriptions. MS COCO has made major contributions to computer vision research by generating cutting-edge models for a variety of visual comprehension problems.

3. **CUB dataset (Caltech-UCSD Birds-200-2011):** The dataset is commonly utilised in computer vision for fine-grained bird species detection. It includes 200 bird species and 11,788 photos in total. Each image in the collection has bounding boxes, part positions, and characteristics labelled on it. The CUB dataset has been used to create and test algorithms for fine-grained classification, attribute prediction, and other bird species recognition tasks.
4. **Oxford-102 Flowers:** The Oxford-102 Flowers dataset is a well-known benchmark dataset for fine-grained flower categorization in the field of computer vision. It has 102 flower categories with a total of 8,189 photos. Each photograph is labelled with the flower species it depicts. The dataset contains a wide variety of floral photos, allowing researchers to create and test algorithms for flower detection, classification, and other tasks. It has been frequently employed in the research of fine-grained visual categorization and the improvement of algorithms in this domain.
5. **KTH Action Recognition:** The KTH Action Recognition dataset is a popular benchmark dataset for recognising human actions in videos. It comprises of six separate films of people walking, jogging, running, boxing, handwaving, and clapping. The collection includes numerous sequences for each activity done by many people and captured from various perspectives. It is a typical dataset for assessing and developing action detection systems, such as those based on motion analysis, spatio-temporal characteristics, and deep learning approaches.
6. **UCF Sports:** The UCF Sports activity dataset is a wellknown benchmark dataset for recognising activity in sports videos. It is a broad collection of videos that capture numerous athletic activities such as basketball, soccer, diving, horseback riding, and more. The dataset provides a diverse variety of action classes captured from various perspectives and under variable settings. It's frequently used for testing and refining action recognition algorithms, allowing academics to progress the field of sports action analysis and video comprehension.

Table 1
Dataset Information

Dataset Name	Dataset Size
YFCC 100M	15 GB
MS-COCO	25 GB
CUB Dataset	1.1 GB
Oxford-102	0.32 GB
KTH Action Recognition	2.2 GB
UCF Sports	1.7 GB

4. Architecture

4.1. Cogview

The tokenizer of CogView, a text-to-image synthesis model, is a vector-quantized variational autoencoder, or VQ-VAE. The model architecture is as follows: The text encoder reads a text caption and generates a sequence of latent codes. The image decoder uses the text encoder's latent codes to generate an image. After training the VQ-VAE to reconstruct pictures, a separate language model is utilised to translate user input text to the VQ-VAE's latent space, where image production happens. A mix of supervised and reinforcement learning losses is used to train the model. To match the produced photos to the written descriptions, the supervised loss is utilised. The reinforcement learning loss is used to encourage the model to create aesthetically attractive pictures. A collection of text and picture captions is used to train the system. While the text encoder is trained using the written captions, the image decoder is trained using the images. The model is trained with the Adam optimizer, which has a 3e-4 learning rate. The CogView model has been demonstrated to be capable of producing realistic pictures from text descriptions. The model was tested on a range of datasets and found to be competitive with existing text-to-image creation techniques. details of the CogView architecture:

- The text encoder is a one-way Transformer that produces a series of latent codes after receiving a text caption as input.
- The text encoder's latent codes are used by the image decoder, a convolutional neural network, to create an image.
- A dataset including 1.56 million Chinese text-image pairings is used to train the algorithm.
- The model is trained for 144,000 steps.
- The learning rate is decayed using a cosine annealing schedule.
- The batch size is 6,144.
- The Adam optimizer is used with a learning rate of 3e-4.
- The model is trained on a mix of 16-bit and 32-bit precision.
- The model uses a technique called Precision Bottleneck Relaxation (PB-Relax) to stabilize training.
- The model uses a technique called Sandwich Layernorm to improve the stability of training.

4.2. dVAE (disentangled Variational Autoencoder)

dVAE (disentangled Variational Autoencoder) is a textto-image synthesis model that generates pictures from text descriptions using a disentangled latent space. The following is the model architecture: The text encoder takes a text caption as input and produces a sequence of latent codes. The image decoder takes the latent codes from the text encoder and produces an image. Because the dVAE's latent space is disentangled, the latent codes reflect distinct features of the picture. This enables the model to provide more realistic and varied visuals. A mix of supervised and reinforcement learning losses is used to train the model. To match the produced photos to the written descriptions, the supervised loss is utilised. The reinforcement learning loss is

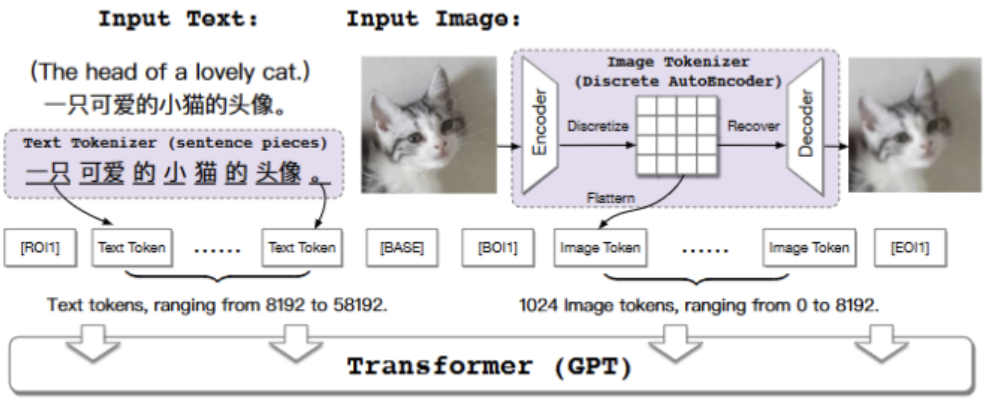

Figure 1: Cogview architecture.

used to encourage the model to create aesthetically attractive pictures. The algorithm is trained on a dataset of picture captions and text captions. The photos are utilised to train the image decoder, while the text captions are used to train the text encoder. The Adam optimizer is used to train the model, which has a learning rate of 3e-4. The dVAE model has been demonstrated to be capable of producing realistic visuals from text descriptions. The model was tested on a range of datasets and found to be competitive with existing text-to-image creation techniques [7]. Here are some of the details of the dVAE architecture:

- The text encoder is a bidirectional LSTM that takes a text caption as input and produces a sequence of latent codes.
- The image decoder is a convolutional neural network that takes the latent codes from the text encoder and produces an image.
- The latent space of the dVAE is disentangled into three factors of variation: pose, shape, and appearance.
- The model is trained on a dataset of 100,000 text-image pairs.
- The model is trained for 100 epochs.
- The learning rate is decayed using a cosine annealing schedule.
- The batch size is 64.
- The Adam optimizer is used with a learning rate of 3e-4.
- Here are some of the specific training details of the dVAE model.
- The model uses a technique called Wasserstein loss to improve the stability of training.
- The model uses a technique called KL annealing to gradually increase the importance of the KL divergence loss during training.

4.3. Multi-Stage AttnGAN

A text-to-image synthesis model called AttGAN (Attention GAN) [8] employs attention to regulate the creation of pictures from text descriptions. The model architecture is as follows: The text encoder takes a text caption as input and produces a sequence of latent codes. The

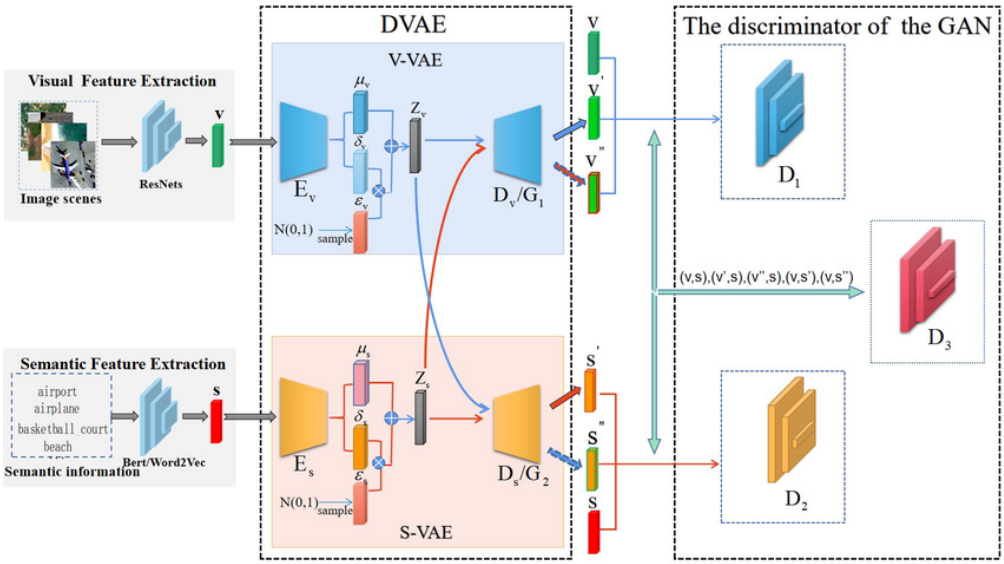

Figure 2: dvae architecture.

image generator takes the latent codes from the text encoder and produces an image. The image discriminator takes an image as input and produces a probability that the image is real or fake. When producing the picture, the attention mechanism allows the image generator to focus on select sections of the written description. As a result, the model may provide visuals that are more compatible with the written description. A mix of adversarial and supervised losses is used to train the model. The adversarial loss is used to teach the image generator and discriminator to compete. To match the produced photos to the written descriptions, the supervised loss is utilised. The algorithm is trained on a dataset of picture captions and text captions. The pictures are utilised to train the image discriminator, while the text captions are used to train the text encoder and image generator. The Adam optimizer is used to train the model, which has a learning rate of 3e-4. The AttGAN model has been demonstrated to be capable of producing realistic pictures from text descriptions. The model was tested on a range of datasets and found to be competitive with existing text-to-image creation techniques [9]. details of the AttGAN architecture:

- The text encoder is a bidirectional LSTM that takes a text caption as input and produces a sequence of latent codes.
- The image generator is a convolutional neural network that takes the latent codes from the text encoder and produces an image.
- The image discriminator is a convolutional neural network that takes an image as input and produces a probability that the image is real or fake.
- The image discriminator is a convolutional neural network that takes an image as input and produces a probability that the image is real or fake.

Here are some specific training details of the AttGAN model:

- The model is trained for 100 epochs.
- The batch size is 64.
- The Adam optimizer is used with a learning rate of 3e-4.

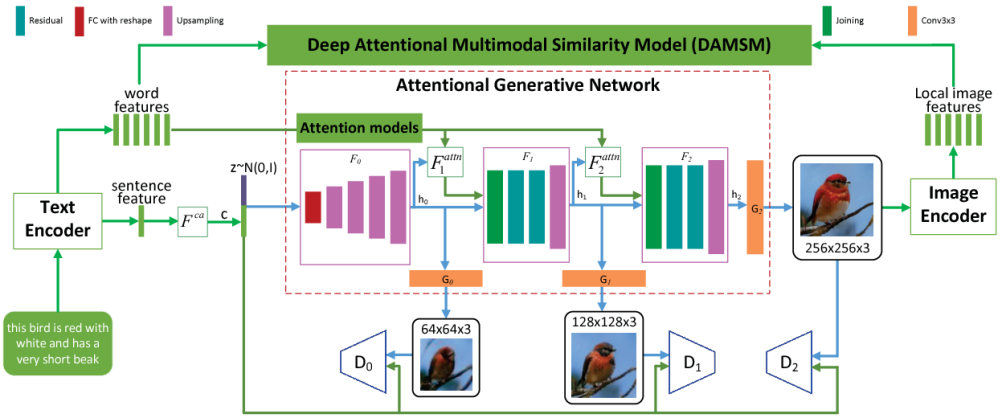

Figure 3: AttGAN architecture.

4.4. CycleGAN + BERT

In this work, the original cycle GAN strategy and the attention GAN technique are combined. With the addition of subtitles and an RNN trained on visual attributes, the mode is enhanced. By converting images back into text, the Semantic Text Regeneration and Alignment Module (STREAM) makes sure that the pictures contain the latent data needed to recreate the original captions. Furthermore, pre-trained BERT encoding transformers are employed in place of standard word embeddings. Constructed from deep, pre-trained language models, these transformers have shown promise in numerous natural language processing applications and aid in the enhancement of the CycleGAN architecture.A text caption is fed into the bidirectional LSTM text encoder, which outputs a series of latent codes.

- The image discriminator is a convolutional neural network that takes an image as input and produces a probability that the image is real or fake.
- The BERT model [10] is a Transformer-based model that takes the text caption as input and produces a sequence of hidden states.
- The image generator is a convolutional neural network that uses the latent codes from the text encoder to create an image.
- The model uses a technique called Wasserstein loss to improve the stability of training.
- The model is trained on a dataset of 500,000 text-image pairs.
- The training goes on for 100 epochs and the learning rate is decayed using a cosine annealing schedule, with a size of 64 for each batch.
- The Adam optimizer is used with a learning rate of 3e-4.

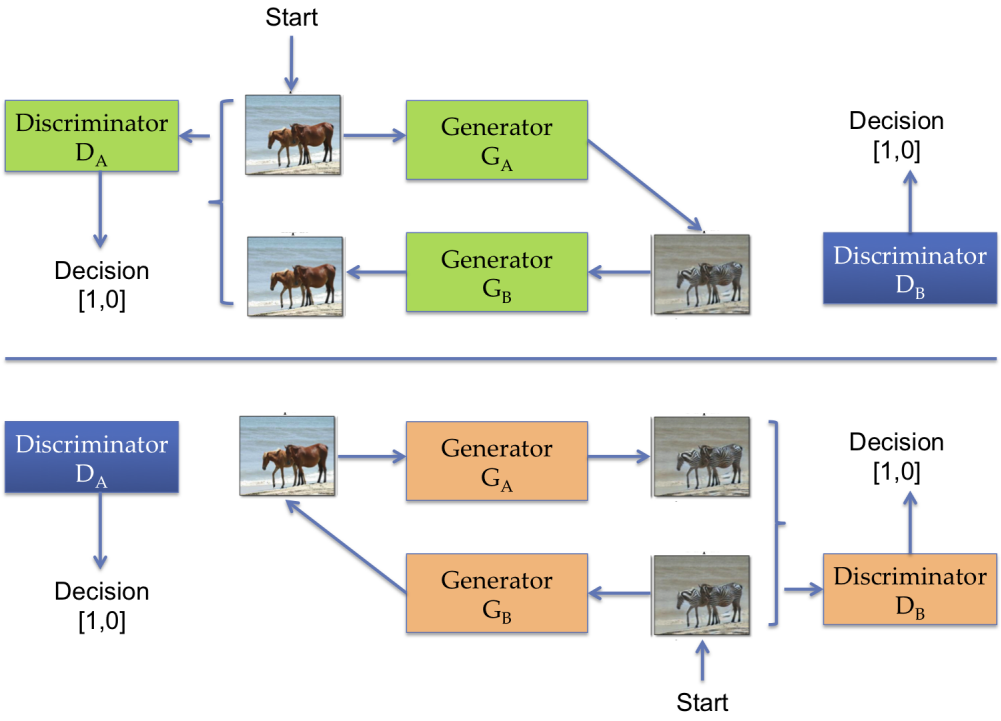

Figure 4: cycle GAN+BERT architecture.

4.5. DF-GAN

TThe DF-GAN [11] is made up of the generator, discriminator, and pre-trained text encoder. To guarantee the diversity of the pictures it generates, the generator needs two inputs: a phrase vector encoded by a text encoder and a noise vector sampled from a Gaussian distribution. The noise vector is first turned into a completely connected layer. The properties of the image are then upsampled using a sequence of UP-Blocks. An upsample layer, a residual block, and DF-Blocks make up the UPBlock, which combines the text and picture attributes throughout the image manufacturing process. Finally, an image feature is converted into an image using a convolution layer. Images are converted into attributes by the discriminator using a sequence of DownBlocks. After that, a copy of the vector phrase is mixed with the properties of the photo. To evaluate the visual realism and semantic coherence of the inputs, an adversarial loss will be anticipated. The discriminator aids the generator in producing pictures of superior quality and textimage semantic coherence by differentiating synthetic images from authentic examples. The bidirectional long short-term memory (LSTM) text encoder extracts semantic vectors from the text description. We employ AttnGAN's pre-trained model directly. Image scaling and normalisation are two ways to preprocess the image data.

- Images undergo resizing and normalizing the images.
- Text undergoes tokenisation followed by verctorisation.
- Text undergoes tokenisation followed by verctorisation.

- Train the image encoder using a dataset of real images.
- Update the image encoder's weights using a suitable loss function (e.g., Mean Squared Error).
- Freeze the generator and discriminator.
- Train the text encoder using the vectorized text descriptions.
- Update the text encoder's weights using a suitable loss function.
- Unfreeze the generator, discriminator, and encoders.
- Generate fake images by sampling from random noise and text features.
- The discriminator part of the GAN is trained to distinguish between real and fake images.
- The generative part of the GAN is trained to fool the discriminator by generating realistic images.

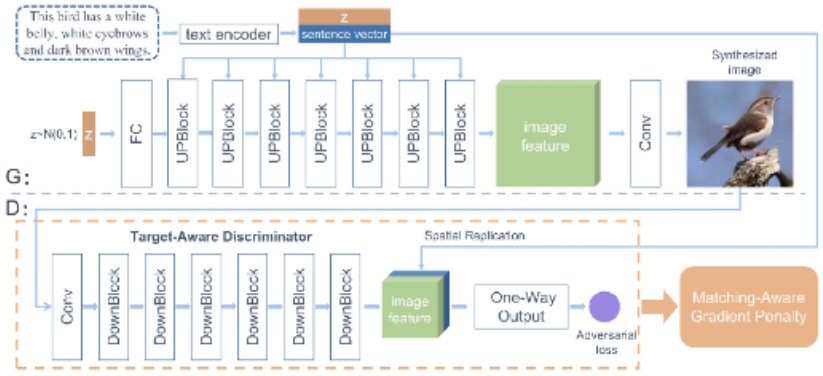

Figure 5: df-GAN architecture.

4.6. MirrorGAN

The figure above depicts the MirrorGAN implementation [12], which includes a mirror structure that combines T2I (textto-image) and I2T (image-to-text) functions. MirrorGAN's fundamental idea is to use the concept of redescription to train T2I generation. MirrorGAN then reproduces the image's description, successfully matching the created image's underlying semantics with the given text description. The MirrorGAN model is made up of three major components: STEM, GLAM, and STREAM. Each module performs a distinct task in the model's overall operation.Pretrain a text encoder network using a large-scale text dataset (e.g., text corpus).

- Pretrain a text encoder network using a large-scale text dataset (e.g., text corpus).
- The text description is fed into this network, which then encodes it into a fixed-length feature vector.
- The weights of the text encoder are updated using an appropriate loss function (such as Cross-Entropy Loss).

- The generator is trained to produce realistic images from text features and random noise.
- The discriminator is trained to discern between generated and real images.
- The discriminator is trained to discern between generated and real images.
- Turn off the discriminator and generator.
- Update the text encoder's weights using an appropriate loss function (such as Triplet Loss) after training it with the dataset's text descriptions.
- Create false images by sampling from text features and random noise.
- Switch between training the discriminator and generator.

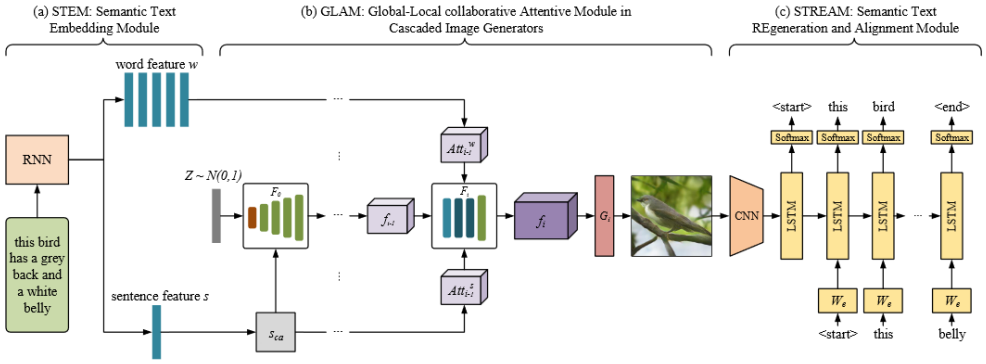

Figure 6: MirrorGAN architecture.

4.7. VQ-SEG - a modified version of the Variational Quantum VAE (VQVAE)

A version of the Variational Quantum VAE (VQVAE) architecture designed for image synthesis and segmentation tasks is called VQ-SEG. The architecture of VQ-SEG consists of several key parts. Using an encoder network, incoming images are first transformed into a representation in a lower-dimensional latent space. In order to capture hierarchical information at many scales, this encoder typically has convolutional layers followed by downsampling techniques like pooling or strided convolutions. The latent space representation is then obtained by the Vector Quantization (VQ) layer. The continuous latent coding is discretely quantized by the VQ layer. It uses a preset codebook that it learned during training to match each latent code to the closest codeword. Computation and storage are facilitated by this distinct form.

The VQ-SEG picture segmentation process now has an additional branch thanks to modifications made to the VQVAE architecture. This branch produces segmentation masks, pixel by pixel, that show the class labels of different regions inside the input image. VQSEG incorporates an extra decoder network to enable picture segmentation. This decoder produces pixel-wise predictions for every class label using quantized latent codes. Often, the decoder network consists of upsampling or transposed convolutions to improve the feature maps' spatial resolution. VQ-SEG integrates segmentation and reconstruction losses during training. The reconstruction loss promotes output images that are similar to the original input images, but the segmentation loss penalises differences between expected segmentation masks and ground

truth masks. Typically, these losses are determined using pixel-by-pixel comparisons, such as mean squared error or cross-entropy loss. In essence, VQ-SEG is a modified VQVAE architecture that combines discrete quantization with segmentation branches to add image segmentation capabilities.

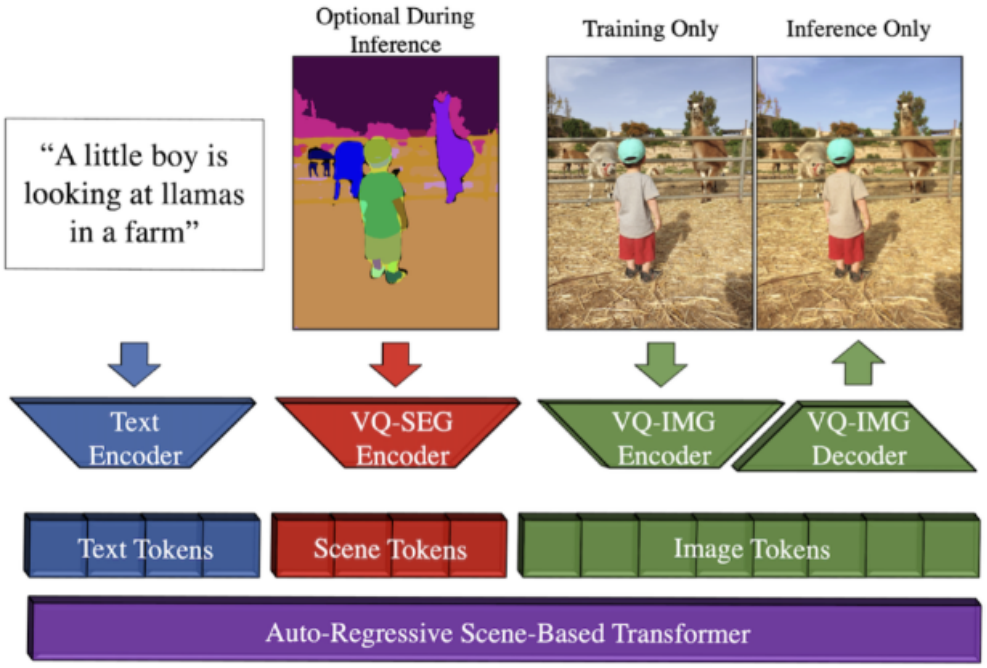

Figure 7: VQVAE: The architecture of the scene-based method: Images are created from input text with optional layout. Transformer creates tokens that networks then encode and decode.

4.8. StackGAN + Fine-tuned BERT Text Encoding Models

Using the BERT model and the StackGAN architecture, realistic visuals are produced from textual descriptions. There are two phases to the architecture: Stage 1 entails turning text into low-resolution images, and Stage 2 concentrates on enhancing those images into high-resolution versions. The target dataset is used to fine-tune a pretrained BERT model in the BERT-based text embedding process. The BERT model can now efficiently comprehend and reflect the semantics of the provided textual descriptions thanks to this fine-tuning. The first stage of the StackGAN model seeks to produce low-resolution pictures that roughly depict the colour and shape mentioned in the textual descriptions. The BERT-based text embedding vector and a random noise vector are sent into the generating network. It generates an image with low resolution that matches the written description. The generated low-resolution images are compared to the real images, which are dependent on the textual descriptions, by the discriminator network. In order to produce high-resolution photographs with better details, Stage 2 concentrates on enhancing the low-resolution images created in Stage 1. The low-resolution image created in Stage 1 and the BERT-based text embedding vector are inputs used

by the generator network in Stage 2. It generates an upscale picture that corresponds with the provided written explanation. The generator's high-resolution images are compared to the actual high-resolution images conditioned on the textual descriptions by the discriminator network in Stage 2. Mini-batches and iterations are used in the StackGAN with BERT training method. Adversarial training is used to update the discriminator and generator networks' parameters. Carefully chosen learning rates for the discriminator and generator guarantee successful convergence during training. The model seeks to produce realistic images [13] that are coherent with the given textual descriptions by combining the power of BERT-based text embeddings with the hierarchical image creation approach of StackGAN.

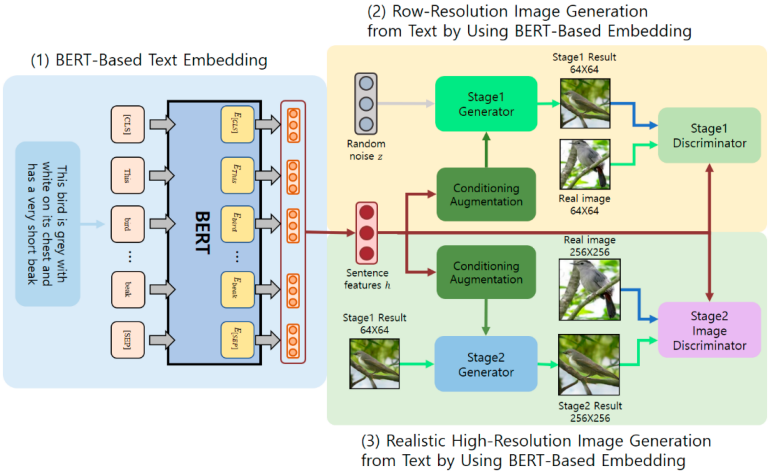

Figure 8: StackGAN + Fine-tuned BERT architecture.

4.9. DALLE-2

Diffusion models are used by the DallE-2 architecture to produce high-resolution images conditioned on optional text descriptions and CLIP image embeddings. CLIP embeddings are projected into and added to the current timestep embedding in this enhanced architecture. In addition, four additional context tokens are projected from CLIP embeddings and concatenated with the GLIDE text encoder's output sequence. It is discovered that the text conditioning pathway provides minimal assistance in this area, despite its attempt to capture natural language elements that CLIP might miss. In order to improve sample quality, training involves randomly removing the text caption 50% of the time and randomly setting CLIP embeddings to zero or applying learnt embeddings 10% of the time. Guidance on conditioning information is also used. It takes two trained diffusion upsampler models to produce high-resolution images. While the second upsampler further upsamples the photos to 1024×1024 resolution, the first upsampler concentrates on boosting the resolution from 64×64 to 256×256. By employing methods like Gaussian blur and varied BSR degradation to slightly contaminate the conditioned images during training, the robustness of the upsamplers is increased. The Dalle2 architecture does not include attention layers; it only uses spatial convolutionals. The model is applied immediately at the

target resolution during inference, demonstrating its capacity to generalise to higher resolutions without the requirement for extra conditioning on the text caption. The upsamplers use the unconditional ADMNets technique and are not conditioned on the caption. To summarise, the GLIDE text encoder, CLIP image embeddings, and diffusion models are used in the Dalle2 architecture to produce high-resolution images. The resulting images' quality and resilience are enhanced by the processes of conditioning and upsampling.

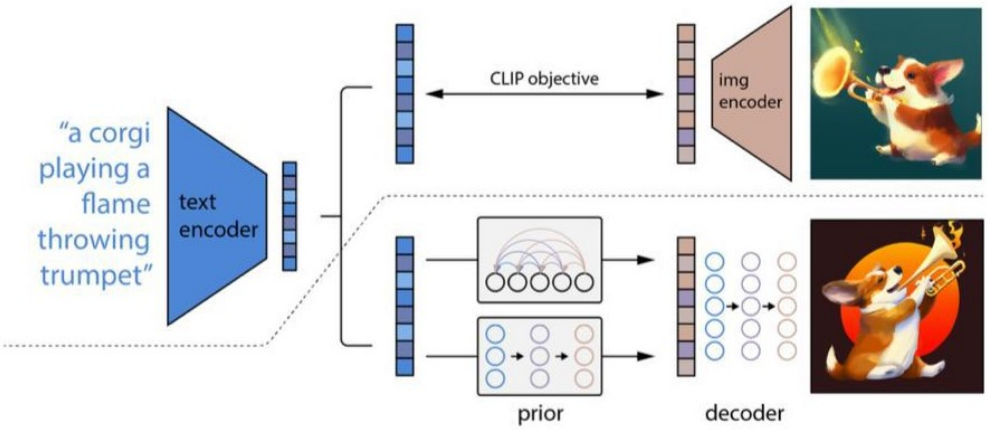

Figure 9: DALLE 2 architecture.

4.10. LSTM+GAN

1. The LSTM (Long Short-Term Memory) model [14], is known for its ability to capture long-range dependencies in data. A single LSTM memory cell, which consists of input gates (it), forget gates (ft), output gates (ot), cell state (ct), and cell input activation vectors. The LSTM model utilizes composite functions to calculate these components based on input and previous hidden states. The model employs the logistic sigmoid function and the hyperbolic tangent function to process the inputs. The original LSTM algorithm used an approximate gradient calculation, but this paper adopts backpropagation through time for gradient calculation. However, training with the full gradient can lead to large derivative values. The LSTM unit receives inputs from external sources at each time step and updates its internal cell state and hidden state based on these inputs and previous states.
2. Recurrent neural networks (RNNs) composed of LSTM (Long Short-Term Memory) units are used in the LSTM Autoencoder Model, an unsupervised learning model. An encoder LSTM and a decoder LSTM are the two RNNs that make up the model. An image patch or set of features is a sequence of vectors that are fed into the model. Following the processing of this input sequence by the encoder LSTM, the decoder LSTM assumes control after all inputs have been read. A prediction for the target sequence—which is the input sequence in reverse order—is produced by the decoder LSTM. Both conditioned and unconditioned decoders are possible. The final created output frame is fed into a conditional decoder, but it is not fed into an unconditioned decoder.

3. The Future Predictor Model shares the same design as the Autoencoder Model, with the key difference lying in the decoder LSTM. While the Autoencoder Model predicts the target sequence that matches the input sequence, the Future Predictor Model goes a step further and predicts frames of the video that come after the input sequence. Essentially, this model is designed to forecast a longer sequence into the future, extending beyond the input sequence.

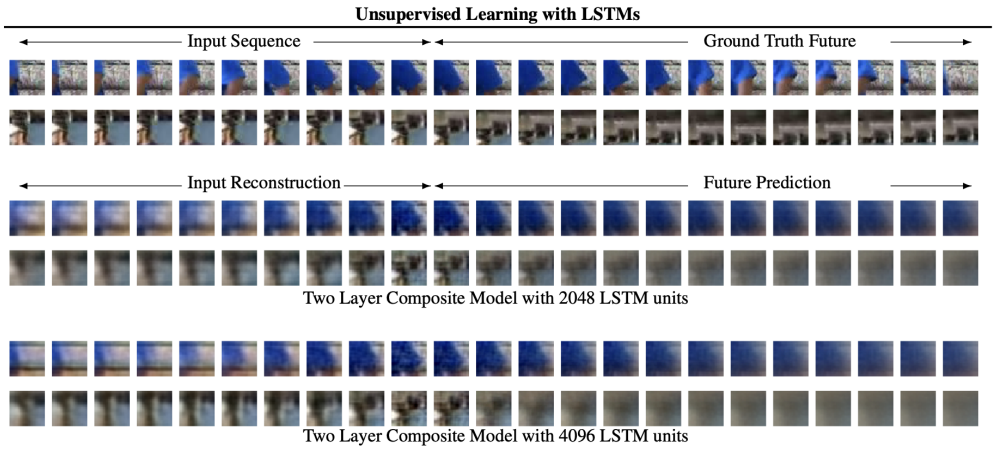

Figure 10: lstm+GAN architecture: The Composite Model forecasts the future of natural image patches. In the first two rows are ground truth sequences. It uses 16 input frames and displays the most recent 10. The true future is shown in the next 13 frames. Here are the predicted and reconstructed frames for two model examples.

Table 2
PERFORMANCE METRICS OF DIFFERENT MODELS

Model Name	Inception Score	Frechet Inception Distance
CogView	32.2	23.6
Discrete Variational Autoencoder	23.6	30
AttentionGAN	4.58	19
GAN	17	23
LSTM + GAN	16	21
VQ-VAE	18.2	23.6
DF-GAN	5.10	19.32
MirrorGAN	4.54	20
StackGAN + BERT	4.44	37.7
CycleGAN + BERT	6	28

5. Comparative Analysis

This section contains the analysis and findings from our investigation, which assessed the effectiveness of the eleven GAN and auto encoder models created for text-to-image conversion. One of the challenges encountered in the CogView framework is the slow generation process inherent to autoregressive models, as images are being generated token-bytoken. Additionally, blurriness is introduced as a substantial restriction by the use of VQVAE. When discretizing continuous data for use with discrete variational auto-encoders (dVAEs), there are disadvantages including limited expressiveness and possible information loss. Due to a lack of textimage pairs for each category and the inclusion of more abstract captions in the dataset, such as COCO, the multistage Attention-GAN model is constrained. Gaps exist in the ability of Generative Adversarial Networks (GANs) to produce coherent, high-quality images that are in line with the input data. While leveraging unsupervised learning, the LSTM+GAN technique struggles to produce clusters that truly reflect the truth, leading to restricted expressiveness regarding input information. CycleGAN+BERT's performance is hampered by insufficient training time and the absence of hyperparameter adjustment because of time restrictions. Due to its strong sensitivity to hyper-parameters, DF-GAN primarily relies on pre-trained models and lacks diversity in generated data. Basic text embedding techniques have a negative impact on STEM integration and the quality of the outcomes for MirrorGAN. The image quality of VQ-SEG, a modified VQVAE, could be better, however the improvements also cause losses in perceptual knowledge and awareness of a specific region. Further studies are required to construct complex loss functions and efficiently create images from text with little data for StackGAN+fine-tuned BERT text encoding models. Finally, Conditional Adversarial Networks (cGAN) still have difficulties in producing visually and semantically cohesive video sequences from textual descriptions.

6. Performance Metrics

Inception Score(IS) A statistic called the Inception Score is employed to evaluate the calibre and variety of images produced by GANs. It indicates both image quality and diversity by measuring the difference between the average class probabilities across all generated images and the individual class probabilities.

Fréchet Inception Distance(FID) A version of the Inception Distance metric created especially for assessing the effectiveness of picture generating models is called Frenchet Inception Distance. By adding Frechet Distance, ´which gauges how similar two distributions are, it increases the initial Inception Distance. The distributions of feature representations taken from a pre-trained Inceptionv3 model for both actual and generated images are compared using the Frenchet Inception Distance. Better picture generating quality is indicated by a lower Frenchet Inception Distance.

Mean Opinion Score(MOS) The generated photos' quality could be evaluated subjectively using the Mean Opinion Score (MOS). On a numerical scale, human participants would be asked to score the generated images' perceived fidelity or quality. The MOS, which provides

an overall assessment of the image quality, would then be calculated by averaging the ratings given by several people. Higher MOS values correspond to more visually attractive or realistic perceptions of the created images, whereas lower MOS values correspond to lower quality or fidelity. User happiness can be measured and text-to-image conversion model[15] enhancements can be directed via MOS evaluations.

7. Future Research Directions

Converting text to image generation models have come a long way, but there are still a number of areas that need more research and development. This section highlights potential future research directions based on the current state of models and identified areas for improvement

Improved Semantic Understanding: Enhancing the semantic understanding of text is crucial for generating more accurate and contextually relevant images. Future research could focus on incorporating advanced natural language processing techniques, such as pre-trained language models or knowledge graphs, to capture a deeper understanding of text semantics. This could enable models to generate images that align more closely with the intended meaning of the input text. One area that could potentially be beneficial is generating knowledge graphs from text embeddings to improve context and positional understanding greatly.

Increased Resolution and Realism: Though current models have come a long way in producing high-quality photographs, resolution and photo-realism may still use some work. Future research could focus on developing techniques to generate images at higher resolutions, allowing for more detailed and visually appealing results. Additionally, exploring advanced loss functions or perceptual similarity metrics could further enhance the realism of generated images, making them indistinguishable from real photographs.

Fine-grained Control and Manipulation: Current text-to-image models often lack fine-grained control over generated images. Future research could investigate methods to enable precise control and manipulation of image attributes, such as object positions, colors, and styles, based on textual input. This could involve exploring novel conditioning techniques or incorporating additional information during the generation process to produce images that align with specific user requirements.

Handling Ambiguity and Multi-modal Outputs: Textual descriptions often contain ambiguous or subjective elements that can lead to multiple plausible interpretations. Future research could explore methods to handle such ambiguity and generate diverse, multi-modal outputs that capture different interpretations of the same textual input. This could involve incorporating uncertainty estimation techniques, exploring variational approaches, or leveraging adversarial learning to encourage the generation of diverse image outputs. One possible approach is training the image generator module on a combination of parsed output from scene graph and the actual prompt for a more objective understanding leading to potentially reduce ambiguity to a reasonable extent.

Incorporating User Feedback and Interactive Generation: Interactive text-to-image generation systems that incorporate user feedback and preferences hold great potential for enhancing user satisfaction and enabling personalized image generation. Future research could focus on developing models that can adapt and refine their generation process based on user

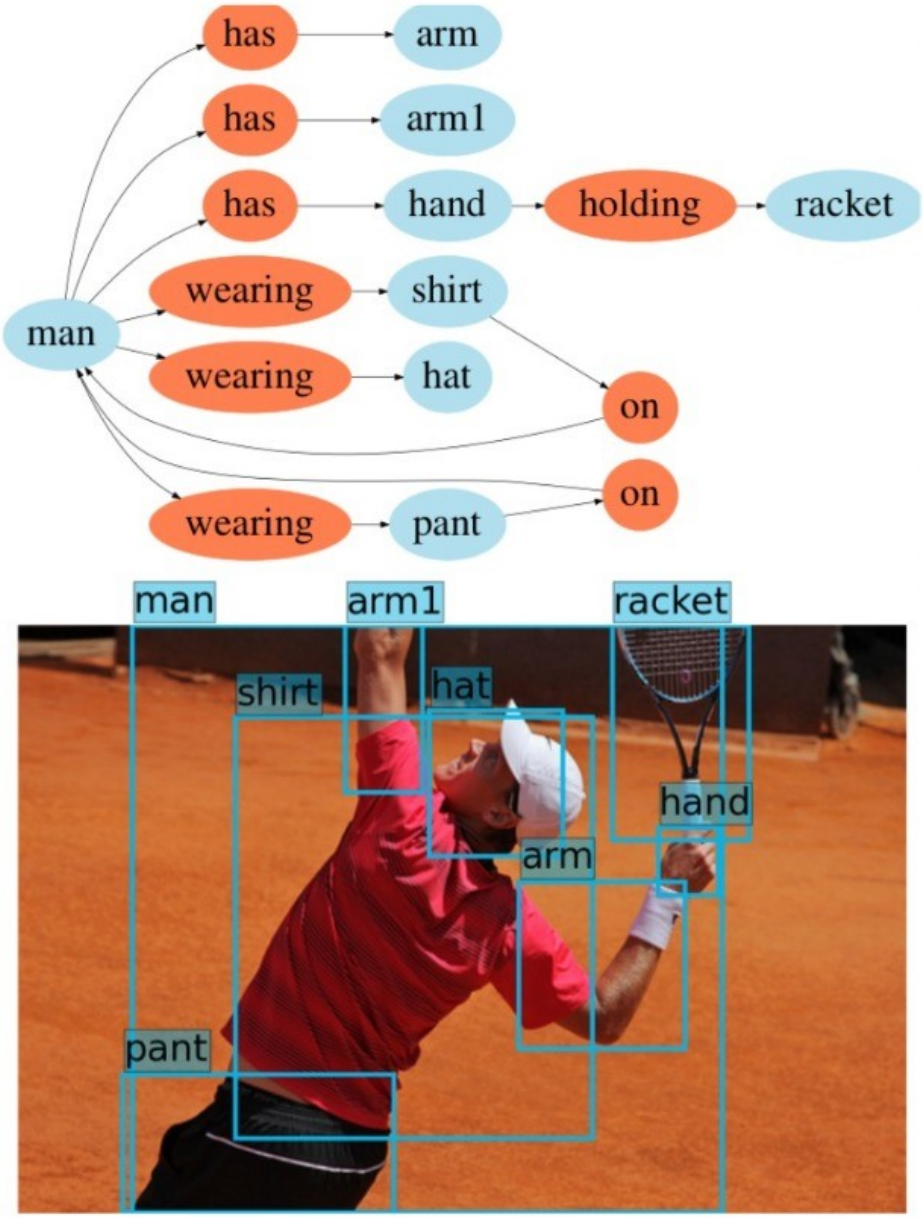

Figure 11: scene graphs giving grater positional knowledge.

interactions, allowing users to provide feedback and guide the image synthesis process in real-time.

Ethical Considerations and Bias Mitigation: As text-to-image generation becomes more prevalent, it is important to address ethical considerations and mitigate potential biases in the generated content. Future research should explore meth- ods to ensure fairness, diversity, and

inclusivity in generated images, avoiding the reinforcement of harmful stereotypes or biases present in the training data. This could involve developing bias detection [16] and mitigation techniques or incorporating fairness constraints during the training process.
These new lines of inquiry could further the field of text-to-image generation and open up new avenues for producing contextually appropriate, high-quality images from textual input. Through the exploration of novel methodologies and resolution of these obstacles, scholars can facilitate the development of more advanced and adaptable text-to-image generation models that have wider applications across diverse fields.

8. Conclusion

In conclusion, our comparative study of 11 text-to-image generation models highlighted StackGAN as the top performer. StackGAN achieved a remarkable inception score of 4.44, indicating its ability to generate visually diverse and high-quality images. Additionally, StackGAN outperformed other models with a FID score of 37.7, demonstrating its superior ability to capture image fidelity and similarity to real images. While other models, such as cogview[17] and dVAE, showcased strengths in specific areas, they fell short in terms of overall performance compared to StackGAN. The models based on GAN architecture, including Multi-Stage AttnGAN, LSTM+GAN, and CycleGAN+BERT, exhibited promising results in capturing global and local image details, but StackGAN surpassed them in terms of both inception and FID scores.Our study also emphasized the impact of dataset selection on model performance. The MSCOCO dataset provided a diverse range of images, contributing to the evaluation and comparison of the models. The outcomes demonstrated that StackGAN could make good use of the dataset, producing better image creation results. These results offer insightful information to the text-to-image generating sector and help practitioners and researchers select models that are suitable for their particular requirements. Future studies can concentrate on developing StackGAN even more and investigating its possible uses in a range of fields, including as virtual reality, multimedia content generation, and computer vision.In conclusion, our comparison analysis shows that StackGAN is the best model for text-to-image creation, doing remarkably well with an origin score of 4.44 and a Fretchet Inception Distance score of 37.7. These outcomes demonstrate the effectiveness of StackGAN as a method for producing realistic and varied graphics from text descriptions.

References

[1] D. Chaudhary, P. Agrawal, V. Madaan, Bank cheque validation using image processing, in: International Conference on Advanced Informatics for Computing Research, https://link.springer.com/chapter/10.1007/978-981-15-0108-1_15, 2019, pp. 148–159.

[2] V. Madaan, K. Sood, P. Agrawal, A. Kumar, C. Gupta, A. Sharma, A. K. Shukla, Solving direction sense based reasoning problems using natural language processing, Machine Learning and Data Science: Fundamentals and Applications (2022) 215–230.

[3] B. Li, X. Qi, T. Lukasiewicz, P. H. S. Torr, *controllable Text-to-Image generation (2021).

[4] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni,

D. Parikh, Sonal Gupta, Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data, 2023.

[5] S. Chauhan, P. Agrawal, V. Madaan, E-gardener: building a plant caretaker robot using computer vision, in: 2018 4th International Conference on Computing Sciences (ICCS), IEEE, 2018, pp. 137–142.

[6] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis.* in: Proceedings of the 33rd international conference on machine learning 48 (2016) 1060–1069.

[7] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, DreamBooth: Fine tuning Text-to-Image diffusion models for Subject-Driven generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[8] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, Attngan: Fine-grained text to image generation with attentional generative adversarial networks, proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2018) 1316–1324.

[9] M. Tao, H. Tang, F. Wu, X. Jing, B. Bao, C. Xu, Df-Gan, A Simple and Effective Baseline for Text-to-Image Synthesis.* arXiv preprint, 2022.

[10] T. Tsue, S. Sen, J. Li, Cycle Text-To-Image GAN with BERT (2020).

[11] M. Tao, H. Tang, F. Wu, X. Jing, B. Bao, C. Xu, DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis, 2022.

[12] T. Qiao, J. Zhang, D. Xu, D. Tao, MirrorGAN: Learning Text-to-image Generation by Redescription. arXiv preprint, 2019.

[13] S. Na, M. Do, K. Yu, J. Kim, Realistic image generation from text by using BERT-Based embedding, Electronics 11 (2022).

[14] N. Srivastava, E. Mansimov, R. Salakhutdinov, *Unsupervised Learning of Video Representations using LSTMs, 2015.

[15] O. Gafni, Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors, 2022.

[16] A. Zehe, L. Konle, L. K. Dumpelmann, E. Gius, A. Hotho, F. Jannidis, L. Kaufmann, M. Krug, F. Puppe, N. Reiter, A. Schreiber, N. Wiedmer, Detecting Scenes in Fiction: A New Segmentation Task,* in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2022.

[17] M. Ding, CogView: Mastering Text-to-Image generation via transformers (2021).