Exploring Text-to-Image Translation with GANs: Bridging Linguistic and Visual Data

1st Devika E S
Computer Science and Engineering
Lovely Professional University
devikaes17@gmail.com

2nd Kotakonda Jhansi Rani
Computer Science and Engineering
Lovely Professional University
jansiranikotakonda@gmail.com

3rd Enjula Uchoi
Computer Science and Engineering(Teacher)
Lovely Professional University
enjula.29634@lpu.co.in

***Abstract*—Generative Adversarial Networks (GANs) changed the way pictures are combined, allowing practical images to be generated from noise. In this work, we extend the GAN framework by conditioning image generation on textual descriptions. Our depicted show, a text-conditional GAN, uses manifest drawings in combination with noise to generate structurally accurate images. It gets ready both commotion vectors and content embeddings signifies prepare for generator while discriminator qualifies the realness of produced pictures perceiving them from real ones. Through extensive experimentation and training on a dataset of images containing distinct text, we demonstrate that the model is capable of generating realistic images given input description. We also discuss problems faced during training, including mode collapse and insecurity.**

***Index Terms*—GAN, Discriminator, Convolutional Neural Networks , Generator,Python**

I. Introduction

One of the most prominent advances in artificial intelligence has arguably been really good fakes generated via Generative Adversarial Networks (GANs) which dramatically changed how we approach generative models, especially when working with images. This novel architecture in GANs consists of two interlocked neural net-works — the generator and discriminator — which are trained through an adversarial approach to generate realistic synthesized samples that mimic genuine examples. In recent times, Conditional GANs (cGAN) have become popular among the deep learning community for their ability to generate data based on certain input conditions.

This is a project which produces images for a given textual descriptions using Conditional GAN. Best described as an example of the task is generate images from text descriptions, which requires understanding semantics in a piece of a given format.

In this cGAN model, The Generator uses a random noise vector as other GANs and present it with the tokenized description of text such that there will be more realistic images for given captions.The current model also uses deep learning methods such as Convolutional Neural Network (CNN) for image processing and generation, machine processed using Natural Language Processing technology converting text data to a format which is usable by the generator network. The project explores the generative adversarial training process, where the generator learns to create more realistic images while the discriminator improves its ability to distinguish between real and generated images.Adding to that exploring the GAN architecture, this project focus on data preprocessing techniques, such as resizing and normalizing images, tokenizing text data, and conditioning the generator on the provided text.

One of the most difficult problems in computer vision and machine learning is to convert text into images or create images from text descriptions. The potential to generate original content from textual input can be seen by the model's ability to produce a variety of visually coherent images that matches the textual attributes supplied by using the effectiveness of textual data and the power of GANs.Recent years have seen more advancements in training. When using natural language and automatic image generation, users can describe visual elements with visually rich text descriptions. The most effective way to communicate and gather information is through visual content, like pictures.

Provide these descriptions. Best alignment The visual content corresponding to the text is determined.More accurate and easier to understand than written text. Synthesizing text into images means using computers. A method to convert human-written text descriptions (sentences or keywords) into their visual equivalents. Using correlation analysis of words and images, Controlled synthesis methods. New unsupervised methods The result is particularly deep generative models. The latest developments in deep learning. These models can be used to generate intelligent visual images. A properly trained neural network provides textual hints to the generative image model. The images are generated using textual descriptions.

II. Traditional Methods for Text-To-Image Synthesis

These systems improved the field of text-to-image synthesis through the changing of text into either static or dynamic visual representations.By focusing on NLU, early stages at text-to-image translation aimed to stop the gap between human and machine comprehension. The Story Imagining

Engine is one system that is tasked to improve narratives with related pictures. finding similarity , picking the images and developing the story , and human feedback based image element retrieval are the steps in the process. Zakraoui et al. dealt with the risks and difficulties that earlier models faced in their analysis of several popular text-to-picture tools and systems.

Conversion of text to picture is another model made to improve communication by using textual data as input to create good appealing visuals. It also used labeling semantics and the idea of picturability, which evaluated the chance of matching words to images, in opposite to more straightforward keyword-based methods. Adding to that, the Word2Image platform deals with on text-to-image generation by assembling images from the Flickr platform using visual clustering, semantic clustering, and correlation analysis.

Similarly the multi-modal system CONFUCIUS converts sentences with action verbs into speech-synchronized animations by acting as a text-to-animation converter. creating text to scene method is provided by WordsEye in the idea of more complex systems, which naturally creates 3D static scenes that matches the textual input. It creates its scenes with a language analyzer and a visualizer. Another model that improved messaging is Chat With Illustration (CWI), which automatically pairs text messages with visual content to give users more interesting and effective interactions. to produce images in the Russian language The Utkus system, which uses single modules for natural language analysis, staging, and rendering, was created. By using techniques like language processing, knowledge base building, and scene generation, the Vishit method also plans to visualize Hindi text.

A system was developed mainly for mobile devices for arabic, aiming at creating instances for Arabic stories, with a focus on education for Arab children through interactive and engaging methods. The Illustrate It! conceptual graph matching were used to solve and to create multimedia mobile learning tools for the Arabic language,mixing text and visuals to enhance learning experiences. Foundation for further advancements was made through these early discovery in text-to-image technologies as well as the challenges that need to be addressed for improved efficiency, , showcasing the growing changes in applications and languages, accuracy, and accessibility.

III. New Methods for Text-to-Image Synthesis

Scientists have made better changes in the field of generating content artificially across various domains in recent times, including text, images, and audio. Allowing machine learning to produce high-quality, multi-modal outputs was made possible by Generative models which have developed as powerful tools in this basis. The process of creating fake or man made instances that closely similar to those in an original dataset is known as generative modeling. Four major types of generative models have been developed to tackle these challenges:

Flow-based models and Diffusion models.

1) Goodfellow et al. in 2014 introduced **Generative Adversarial Networks (GANs)** which was the most well-known works ever. The generative adversial network involves of two different neural networks which are a generator and a discriminator.The discriminator tries on distinguishing real images from fake ones. GANs have been used for various applications, such as generating faces which are lifelike and even interchanging between images and text increasing image resolution. The generator creates fake or unreal images from noisy ones .This generative adversarial process goes on till the discriminator cannot tell the difference between real and fake images, leading to the generation of more realistic images.

2) **Variational Autoencoders (VAEs)** provide the facility to the redevelopment of input data when the content is lost and a probability based framework to encode data into a low dimensional latent space, . Its not like traditional autoencoders, VAEs captures a distribution over the latent space, which makes them particularly opted for tasks like data generation and representation learning.

3) **Flow-based generative models** focus on learning different encoders and decoders through transformations guided by neural networks. not like VAEs these models use invertible transformations to encode and decode data, which allows for an exact reconstruction of the input.

4) **Diffusion models** have recently gained recognition as a cutting-edge deep generative approach, particularly in the domain of image synthesis, where they have shown significant improvements over GANs. These models gradually add Gaussian noise to data during the forward diffusion process and learn to reverse this noise in the backward diffusion phase, recovering the original data. Diffusion models are good working in tasks like super-resolution, in-painting, and image editing . they are based on principles from non-equilibrium thermodynamics.

1) : These models each of them represents an identical approach to the difficulties of generative modeling, with diffusion models currently leading the way in terms of their generative capabilities.

IV. Literature Review

A. GANs in Image Generation

Generative Ill-disposed Systems, presented by Goodfellow et al. (2014), have been broadly received in the space of picture era due to their capacity to create reasonable pictures. The essential thought includes two systems: a generator that makes pictures and a discriminator that endeavors to separate between genuine and produced pictures. This antagonistic preparing leads to the generator progressing over time. [1]

B. Conditional GANs

The concept of Conditional GANs amplifies conventional GANs by conditioning both the generator and discriminator on extra data such as course names, empowering control over the era handle. More later work joins literary information as the conditioning variable, empowering the era of pictures based on composed descriptions. [2]

C. Text-to-Image Synthesis

Reed et al. (2016) showed the potential of text-to-image GANs, where it demonstrate creates pictures based on literary explanations. The challenge lies in guaranteeing or ensuring that the created picture precisely reflects the semantics or details of the input content, which includes complex connections between the content and picture domains. [3]

D. Scaling up gans for text-to-image synthesis

Discovery of using StyleGAN shows, when its capacity is increased the architecture becomes unstable.to overcome this GigaGAN is a novel model used ,they are more practical in text-to-image synthesis containing mainly 3 benefits-512px image can be synthesized in 0.13 seconds,creating high resolution image in 3.66 sec like 16 megapixel images,and vector arithmetic and style mixing. [4]

E. Spatial Fusion GAN for Image Synthesis

For achieving realism in geometry and appearance spaces ,the paper introduced SF-GAN that contains both geometry and appearance synthesizer which can be used to convert foreground objects,colors,brightness to background images using guided filters by using mutual references. [5]

F. Text to Photo-Realistic Image Synthesis With Stack-GAN

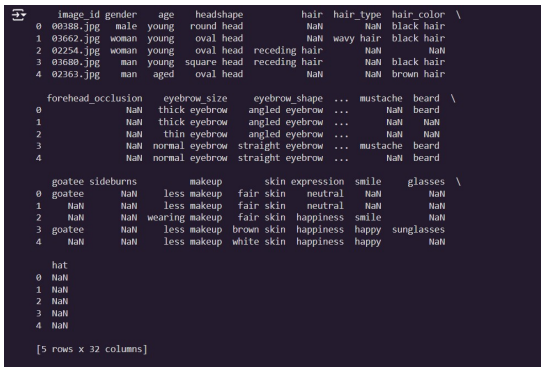
Stack-GAN can generate stage-1 low resolution images by using the object's basic shapes and colors ,for high realistic images stage-2 gan is used where refinement stage is present along with conditioning augmentation which offers smoothness in the latent conditioning manifold. [6]

V. Methodology

Our approach uses a text-conditional GAN, which includes conditioning both the generator and the discriminator on printed depictions. The strategy can be splitted down into the taking after components: dataset planning, content encoding, GAN design, and the preparing procedure.

A. Dataset Features

Datasets play a vital part in improvement and assessment. A demonstrate that changes over generative content into pictures. In the field of generative text-to-image models, utilizing different datasets is exceptionally critical to get exact and reasonable visual comes about. In this area, we will see at different commonly utilized datasets. The most commonly utilized datasets in this field of consider Text-to-image union models:



```
   image_id gender    age    headshape          hair  hair_type  hair_color  \
0  00388.jpg   male  young   round head           NaN        NaN  black hair
1  03662.jpg  woman  young    oval head           NaN  wavy hair  black hair
2  02254.jpg  woman  young    oval head  receding hair        NaN         NaN
3  03680.jpg    man  young  square head  receding hair        NaN  black hair
4  02363.jpg    man   aged    oval head           NaN        NaN  brown hair

  forehead_occlusion    eyebrow_size     eyebrow_shape  ...  mustache  beard  \
0                NaN   thick eyebrow    angled eyebrow  ...       NaN  beard
1                NaN   thick eyebrow    angled eyebrow  ...       NaN    NaN
2                NaN    thin eyebrow    angled eyebrow  ...       NaN    NaN
3                NaN  normal eyebrow  straight eyebrow  ...  mustache  beard
4                NaN  normal eyebrow  straight eyebrow  ...       NaN  beard

  goatee sideburns          makeup        skin expression  smile     glasses  \
0  goatee       NaN     less makeup   fair skin    neutral    NaN         NaN
1     NaN       NaN     less makeup   fair skin    neutral    NaN         NaN
2     NaN       NaN  wearing makeup   fair skin  happiness  smile         NaN
3  goatee       NaN     less makeup  brown skin  happiness  happy  sunglasses
4     NaN       NaN     less makeup  white skin  happiness  happy         NaN

   hat
0  NaN
1  NaN
2  NaN
3  NaN
4  NaN

[5 rows x 32 columns]
```

B.

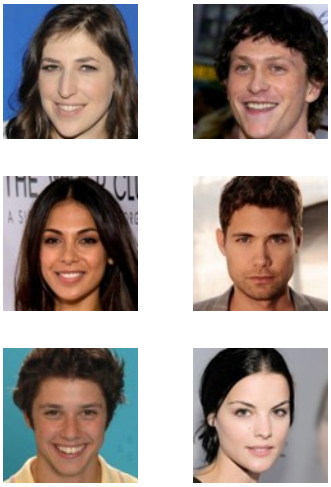
Our dataset comprises of pictures and comparing literary portrayals. The literary portrayals incorporate highlights such as sexual orientation, age, head shape, and hair color. Each depiction is a organized sentence depicting the appearance of the individual in the picture. For preparing, we prepare both the pictures and the literary descriptions. Dataset structure:

- **Image ID**:identification for each image to make it unique
- **Gender**: to check if the person is male or female
- **Age**: to find Age range of the individual.
- **Head Shape**: to find if the person has Round or oval or Square face, etc.
- **Description**: A combination of the above highlights, e.g., "Male, 25, round head, black hair."

Image id	Gender	Age	Head Shape	Hair Color	Description
001	male	25	Round	Black	Male 25 round head
002	female	30	Oval	Brown	Female 30 oval head

C. Data Preprocessing

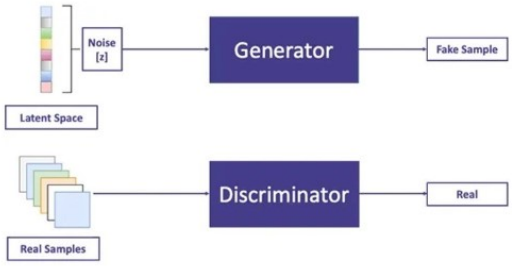
- Text Tokenization:to Clarify the Tokenizer handle, with the rationale of changing portrayals into sequences.
- Text Padding: Discuss why text sequences are padded to a fixed length and how this impacts the training process.



- Image Preprocessing: Expound on how the pictures are resized and normalized (e.g., why scaling the pixel values to [-1, 1] makes a difference with speedier GAN convergence).

D. GAN Model Architecture

- Generator Demonstrate: Give charts for each layer (commotion input, content inserting, convolutional layers, and transposed convolutions). Clarify why each layer was chosen (e.g., why LeakyReLU instep of ReLU, why group normalization, etc.).
- Discriminator Show: Clarify the Conv2D layers and why we diminish the picture measure by half at each layer. Talk about the utilize of sigmoid actuation for parallel classification.



E. Text Encoding

To handle the literary portrayals, we to begin with tokenize and insert the depictions into a fixed-length vector representation. We utilize the Word2Vec or Glove embeddings to change over each word into a vector, taken after by averaging the vectors for each word in the portrayal to get a last content inserting vector yyy. The implanting prepare can be depicted as:

y=1/nEmbed(w)

where w speaks to the i-th word in the portrayal, n is the add up to number of words, and Embed(w) is the word inserting vector for w.

F. GAN Architecture

The GAN architecture consists of two primary components: the generator and the discriminator.

1) ***Generator G(z,y)****:*

- Input: A arbitrary commotion vector zR100 and the content inserting vector yR128.
- Dense layer: The joined vector is passed through a completely associated layer and reshaped into a 32x32 highlight map.
- Concatenation: The clamor vector and the content containing are joined into form into a single vector z=[z,y].
- Transpose Convolutions: The highlight outline is upsampled utilizing transpose convolutions to produce a 128x128 RGB image.

2) ***Discriminator D(x,y)****:*

- Input: A 128x128 picture xxx and the content implanting yyy
- Concatenation: The content inserting is duplicated spatially and concatenated with the picture features.
- Convolutions: The combined picture and content highlights are passed through different convolutional layers.
- Output: A single scalar showing whether the input picture is genuine or fake.

G. Loss Functions

There are 2 model generator and discriminator,which are made using the double-cross entropy misfortune.Both work vice versa like generator points to reduce the misfortune work,while Discriminator points to increase the work.The main aim is to enhance both generator and discriminator in the substituting steps.

1) : Ever since the presentation of GAN–based text-to-image blend in 2014 happened,we can see a good advancement and improvements such as the work of Reed et al,based on profound convolutional GANs.Earlier models had made pictures commonly basing on the lesson names or captions,without bothering about the postures or areas.To address this,the Generative Antagonistic What-Where Arrange (GAWWN) was presented,overcoming the flaws from previous one,like to produce pictures based on drawing like what to draw and where to keep it.So,the main outcome here is GAWWN is aiming to use bounding boxes and key focuses techniques to analyze the protest situations and its locations here the data is mostly visual data.

2) : Stacked Generative Ill-disposed Systems (StackGAN) is used to upgrade the quality of created pictures, presented a two-stage prepare. In the begin with arrange, a low-quality picture is created based on a content explained like a portrayal, and in the moment organize, a high-resolution picture is delivered by capturing more practical and refined and real subtle elements. StackGAN's latest version,

StackGAN++, is often presented a tree-like structure for its generators and discriminators, permitting for multi-scale picture generation and more solid preparing behavior. For more better control, the union handle by utilizing attention-driven components to refine pictures at different stages is progressed by the Attentional Generative Ill-disposed Organize (AttnGAN). AttnGAN centers on key normal a particular form of a language which is peculiar to a specific terms, permitting it to create more exact and point by point picture highlights based on content input.

3) : MirrorGAN is used to take advantage by presenting a text-to-image-to-text engineering that makes sure that the semantic relationship between printed descriptions and produced pictures are always maintained. This design produces pictures and at that point attempt to achieve a goal to reproduce the unique content from those pictures, pointing to maintain coherence over modalities. These pathways highlight the improved scene of GAN-based text-to-image blend, with progressively modern models contributing to higher-quality, more adaptable, and semantically exact picture generation

VI. Challenges

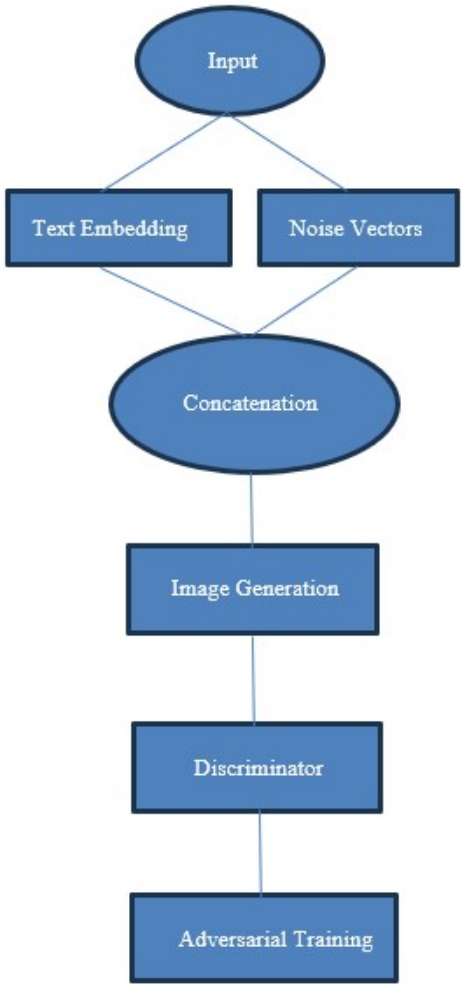
In this work, we utilized a Generative Adversial Network (GAN) for text-to-image blend. Regularly, preparing a GAN requires an large number of training epochs to optimize the performance both the generator and the discriminator. Be that as it may, due to computational difficulties are faced in the utilization of Google Colab, where preparing assets are capped, we prepared the show for as it were 100 epochs instead of the suggested 10,000 epochs.

The project had not been trained or prepared for a adequate number of epochs to capture the fundamental information dispersion The inadequate number of epochs directly cause changes which are affecting the quality of the created pictures, driving to the world of loud and ambiguous noisy pictures.By increasing the number into more epochs the generator would have learned to deliver clearer and more practical pictures by refining the designs and highlights generated from the input captions. This try highlights the significance of drawn out preparing in GAN-based models for accomplishing high-quality results.

In order to proceed further the project needed more computational resources or strategies like using a pre-trained model or changing or moderating the plans in an optimized way to get a chance to long yield by keeping up attainable preparing lengths.

A. *Hardware and Program Constraints*

Generative Adversarial Networks or GANs were utilized for this project for text-to-image union and google colab which is a free site or tool used to run python code is utilized which gives cloud based GPU assets but in a limited manner.Due to



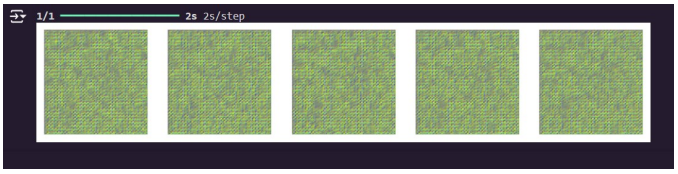
the computational requests of preparing GANs, particularly for text-to-image errands, utilizing a cloud stage was fundamental. The particular GPU accessible amid preparing was a NVIDIA T4 GPU, commonly given in Colab's free tier. The essential computer program environment was based on Python 3.x, utilizing the taking after key libraries:

- **PyTorch**: For executing and preparing the GAN model.
- **Numpy**: For numerical operations.
- **PIL** (Python Imaging Library): For picture processing.
- **H5Py**: For stacking the dataset put away in HDF5 format.
- **TQDM**: For observing advance amid training.

Above this given the setup for GPU get to and computational management, which set the challenges in preparing advanced models like GANs that require wide range of time for coming together.

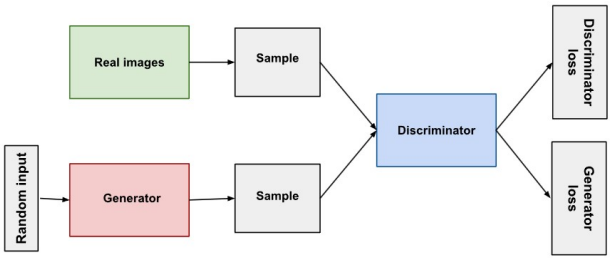
B. *Training Strategy*

The preparing handle taken after standard GAN preparing practises, but due to imperatives forced by the equipment



environment (restricted time on the GPU and constrained computational control), a few alterations had to be made.

- Batch Measure: The bunch measure was set to 64 to adjust memory utilization and GPU handling control. A bigger group estimate would have required more memory, which was not attainable in the Colab environment.
- Number of Ages: The GAN demonstrate was at first prepared for as it were 100 ages due to Colab's GPU time limitations. GAN models, especially for text-to-image amalgamation, regularly require thousands of ages (e.g., 10,000) to deliver important comes about. The constrained number of ages brought about in boisterous yields since the demonstrate didn't have adequate time to learn the mapping between content embeddings and picture generation.
- Learning Rate: A moderately moo learning rate of 0.0002 was utilized with the Adam optimizer to guarantee steady preparing and maintain a strategic distance from mode collapse, which is a common issue in GAN training.



C. *Strategies to Overcome Computational Limits*

- Epoch Decrease: To suit computational limits, less ages were utilized in starting trials, driving to less refined picture outputs.
- Early Halting: This was considered as a potential methodology to stop preparing once the demonstrate accomplished a craved misfortune, but due to the computational constrain, early halting was not altogether actualized in this case.
- Data Enlargement: Strategies like arbitrary flipping and picture revolution were investigated to misleadingly increment the dataset estimate and offer assistance the show generalize way better in spite of the lower number of ages.

VII. Conclusion

We can see that development of GANs and diffusion models have given way to the more advanced models.From this we can say that in the area of converting text to image there was a good progress.We are even seeing the realistic images from normal text description.We can see this happening because as the models are learning from a large sets of data and high quality images.There generating capabilities have been increased.We can conclude this by saying that this model has evolved from basic by learning now it reached to advanced.Here it started learning from big datasets by checking some metrics and it has faced some challenges as well.From that it had overcome from that challenges and opened the doors for new opportunities.But still,it need some improvements,so researchers and some other people should continue to study and improve the models.

VIII. Future Scope of Work

The future scope of work in text to image tool provided various new technologies for enhancement. The biggest area that needs to be improve in Gan architectures such as styleGAN and BigGAN which are able to generate with more accurate and relasitic images.This model can be evaluating in the current frameworks to improve image completeness and accuracy. Likely, the observation of structure like AttnGAN, which prevents the model to work more focus on the given special word as a input, this help the model to provide correct image for a given input.

1) : Moreover, using the more creativity into text embeddings such as BERT or GPT could meaning full to improve to get more clear and represent of text input, which gives the correct image output to the given input. Providing this information to the language models, the text input can provide good output, getting the knowledge about the given input and created images.The incorporation of larger , more large datasets is another crucial and critical step for the future scope . Variant types of dataset like CelebA-HQ or FFHQ,which include more complex and diverse annotations, can focus in generating images that are more normalised and applicable to reallife scenarios.

2) : Unsupervised learning is a technique which mostly algorithms to identify unlabeled data. It's also used in medical fields and in different different fields. By using unsupervised learning, generative ai models can learn from many unlabeled data. Not only focusing on improving technology but also these improvements shows opportunities for various fields like medical field to explain detailed anatomy illustrations and etc... in design field to focus on design hierarchy, font typography and other settings, in advertising field to create quick and message Oriented creative visual for ad sets etc... . We definitely have a risk because it's creating realistic images, risk like misleading information, faking media and manipulation. That's why it's very very important to set up Ethical guidelines and

practices. Hence developer and user should use for transparent purposes.

3) : The main aim of the open-source generative model is to encourage collaboration between researchers and developers. This makes the model easily accessible and simple to handle. Additionally, innovation in open-source generative models has improved and accelerated. This openness in the field speeds up progress and encourages a more important approach to developing technology. Another promising direction is the development of multiple languages and understanding different languages to create text-to-image generation.these ai generative models focuses on services multiple areas audiences so it is also trained to understand different languages and aimed to serve diverse range of audience. this provides equal access for resources and knowledge in all over world.

4) : One of the major advancements is the move to-wards open-source generative models, which encourages collaboration between researchers and developers. By making these models freely accessible, the pace of innovation is accelerated, as re-searchers can build on existing work, test new ap-plications, and improve the performance of cur-rent AI systems. This openness in the field fosters rapid advancements and promotes a more inclusive approach to technological development. Another promising direction is the development of multilingual and cross-lingual models for text-to-image generation. These models aim to break down language barriers, providing a unified foun-dation for understanding and processing multiple languages simultaneously. By doing so, they ena-ble broader access to information and promote linguistic diversity in communication. This could lead to significant improvements in the accessibil-ity of AI tools for users across different linguistic backgrounds.

5) : With Addition to these improved advancements, where improvement mainly focused on reducing the amount of energy and on the computer powers these models needed, focusing on how much time and processing its taking to generate images. However the current best Ai models require lots of computing ppower. they require advanced hardware like expensive graphic card and processors which require lots of energy and also costly that means its not available foreveryone especially like small companies, researchers and individuals has limiting accessibility to experiment. As these models are not available for everyone researchers are focusing on optimising the generative models by adjusting model design pr training method o make them less resource intensive without sacrificing quality and also focusing more on sustainability and accessibility which means affordable, widley available and also eco friendly.

6) : It's not limited to one field, for wide audience it's useful examples fields like. Education, product design and marketing field where visuals are valuable. it has seamless integration from written content with images and making the images from texts quick and straightforward. We use visuals in between long copy to engage audience and to make it easier for them to understand now we can also achieve this by creating illustrations and infographics with these model's. There are particular industries which highly depends on visuals field like graphic design, filmmaking, photography and landing pages or website designers as we expected they are greatly beneficial now with many advancements in Generative models. Creators can be more creative and time saving with the help of these advancements.

7) : The important thing is there's huge potential in upcoming days as still there's so much improvement going on to push technology forward. Developers and researchers are mainly focusing on improving accuracy to match the text provided, as mentioned ongoing improving are going on in future it will focus more on even minor things like tones, shadowing, effects, multiple languages capability for a variety range of audience and also focusing on ethical responsibility by setting some guidelines and safety measures.

IX.

REFERENCES

[1] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[2] Das, Hari Prasanna, et al. "Cdcgen: Cross-domain conditional generation via normalizing flows and adversarial training." arXiv preprint arXiv:2108.11368 (2021).

[3] Reed, Scott, et al. "Generative adversarial text to image synthesis." International conference on machine learning. PMLR, 2016.

[4] Kang, Minguk, et al. "Scaling up gans for text-to-image synthesis." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[5] Zhan, Fangneng, Hongyuan Zhu, and Shijian Lu. "Spatial fusion gan for image synthesis." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[6] D. Joshi, J. Z. Wang, and J. Li, "The story imagin-ing engine—A framework for auto-matic content illustration," ACM Trans. Mixed media Comput., Commun., Appl., vol. 2, no. 1, pp. 68–89, Feb. 2006, doi: 10.1145/1126004.1126008.

[7] Zhang, Han, et al. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.

[8] Wang, Hao, et al. "Cycle-consistent inverse GAN for text-to-image synthesis." Proceedings of the 29th ACM International Conference on Multimedia. 2021.

[9] Mishra, Priyanka, et al. "Text to image synthesis using residual gan." 2020 3rd International conference on emerging technologies in computer engineering: Machine learning and internet of things (ICETCE). IEEE, 2020.

[10] Gragnaniello, Diego, et al. ”Are GAN generated images easy to detect? A critical analysis of the state-of-the-art.” 2021 IEEE international conference on multimedia and expo (ICME). IEEE, 2021.

[11] Wang, Ting-Chun, et al. ”High-resolution image synthesis and semantic manipulation with conditional gans.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[12] Perarnau, Guim, et al. ”Invertible conditional gans for image editing.” arXiv preprint arXiv:1611.06355 (2016).

[13] Thekumparampil, Kiran K., et al. ”Robustness of conditional gans to noisy labels.” Advances in neural information processing systems 31 (2018).