Generating Images from Text Descriptions using Diffusion and CLIP

Sharath Krishna A H
Department of Computer Science and Engineering
PES University
Bengaluru , India
ahsharath@gmail.com

Shrinidhi K J
Department of Computer Science and Engineering
PES University
Bengaluru , India
shrinidhikjkj@gmail.com

Shivani G Itagi
Department of Computer Science and Engineering
PES University
Bengaluru , India
itagi.shivani2002@gmail.com

Dhanvin S
Department of Computer Science and Engineering
PES University
Bengaluru , India
dhanvinsreddy@gmail.com

Dinesh Singh
Associate Professor
Department of Computer Science and Engineering
PES University
Bengaluru , India
dineshs@pes.edu

***Abstract*— In this paper, we introduce a novel approach for generating high-quality images from textual descriptions leveraging a hybrid architecture combining Stable Diffusion UNET and CLIP (Contrastive Language-Image Pre-training). The proposed model aims to bridge the semantic gap between natural language descriptions and visual representations by integrating stable diffusion techniques with the powerful contextual understanding provided by CLIP. The Stable Diffusion UNET architecture is employed to enhance the stability of the generative process, mitigating issues such as mode collapse and improving overall image quality.The CLIP model is integrated into the architecture to facilitate the alignment of textual and visual features, enabling the network to better understand and synthesize images that accurately represent the input textual descriptions.This integration enhances the interpretability and coherence of the generated images, ensuring they align closely with the intended textual prompts.This work contributes to advancing the state-of-the-art in text-to-image synthesis by combining stable diffusion techniques with the contextual understanding of CLIP. The proposed model not only addresses the challenges of mode collapse but also enhances the semantic coherence and visual fidelity of generated images, paving the way for more robust and interpretable text-to-image synthesis systems.**

I. INTRODUCTION

Generating Images from Text Descriptions is a widely researched and evolving field in Machine Learning , Artificial Intelligence , Computer Vision and Natural Language Processing.

Main goal is to create visual representations from the given text inputs which will help the user to visualize and generate images that correspond to the descriptions given.

Earlier Methods relied on predefined rules and correlation between text and image features.However the evolution of deep learning techniques and neural networks helped revolutionize this field. The most important breakthrough was Generative Adversarial Networks which consisted of both Generator and Discriminator . The generator was used to create images from the given text inputs and the discriminator classified the images and evaluated them . Due to the evolution of GAN's , newer techniques such as Stable Diffusion and Transformer models were found to be much more effective for this task.

For the text to image models to be effective , the dataset should be easily available and should be diverse to provide satisfactory results. Dataset should contain captions corresponding to images to help with the training process.

These models are applicable in design and art , helpful in education by creating visual aids. It also has a significant impact in fields such as gaming , entertainment and filmmaking .

With increase in the technology capabilities it's important to keep in track of ethical considerations. The main concern is its potential for misuse such as creating deep fakes or misleading images.

This field has significant scope as it looks towards more sophisticated models that can understand and interpret more complex text descriptions and generate images with much finer details and better quality . The integration of more advanced NLP and improvements in model architectures will continue to push the boundaries of all the possibilities of this field. It provides potential promise in the field of image generation from text description.

II. REVIEW OF LITERATURE

A. CLIP-GEN: Language-Free Training of a Text-to-Image Generator with CLIP

Zihao Wang, Wei Liu, Qian He, Xinglong Wu, Zili Yi, ByteDance Inc.

Massive amounts of paired text-image data are needed for training a text-to-image generator in the general domain, but the cost of collecting this data is prohibitive. In this work, they propose a self-supervised scheme for general text-to-image generation called CLIP-GEN. They extract the language-image priors using a pre-trained CLIP model.

To train the generator, it needs a set of unlabeled images from the public domain. They assert that their approach can even outperform industry standard supervised models like CogView. Early methods that use a convolutional generator to directly generate pixels from the provided text embeddings have demonstrated promising breakthroughs in the generation of images in certain domains. Nevertheless, these methods perform poorly in terms of text-image matching and image quality when used to generate images in the general domain.

Transformer-based text-to-image generators have advanced significantly recently, including DALL-E and CogView. There are two reasons for this advancement. First, by creating discrete representations of images, vector quantized models like VQ-VAE and VQ-GAN allow images to be represented in a manner similar to that of natural language. This makes it possible to train a transformer using cross-modality text-image data within a single framework. Second, large model training—which involves tens or hundreds of billions of parameters—has advanced, greatly utilizing the model's ability to represent cross-modality data in a variety of general domains. Without any paired text-image data, they suggest a method to train a dependable and all-purpose text-to-image generator using a set of unlabeled images and a previously trained CLIP model. This method offers a viable new path for resource-accessible high-fidelity text-to-image generation. Both qualitative and quantitative assessments confirm that this approach maintains text-image matching while outperforming CNN-based and optimization-based text-to-image methods in terms of image quality. The model is even capable of achieving comparable performance to the industry-leading supervised model, CogView, which is trained on enormous quantities of paired data.They use MS-COCO and ImageNet as two datasets to train and assess our methods.

B. *Generating Diverse High-Fidelity Images with VQ-VAE-2*

Ali Razavi, Aäron van den Oord, Oriol Vinyals

The authors imply that this model is used to scale and enhance the autoregressive priors used in VQ-VAE to generate synthetic samples of much higher coherence and fidelity than possible before. It uses feed-forward encoder and decoder networks. The model is able to rival GAN's in terms of quality of images on datasets such as ImageNet and does not face the drawbacks such as mode collapse and lack of diversity.

This model compresses images into a discrete latent space by vector-quantizing intermediate representations of an autoencoder. These representations are over 30x smaller than the original image, but still allow the decoder to reconstruct the images with little distortion. Training and sampling of this generative model over the discrete latent space is also 30x faster than when directly applied to the pixels, allowing us to train on much higher resolution images. Finally, the encoder and decoder used in this work retains the simplicity and speed of the original VQ-VAE, which means that the proposed method is an attractive solution for situations in which fast, low-overhead encoding and decoding of large images are required.

It uses a two stage approach, first stage involves training VQ-VAE to encode images on a discrete latent space and then the second stage uses PixelCNN over the discrete latent space to improve quality of images generated.

Encoder and decoder architectures are kept simple and light-weight as in the original VQ-VAE, with the only difference that this model uses hierarchical multi-scale latent maps for increased resolution. The fidelity of best class conditional samples is competitive with the state of the art Generative Adversarial Networks, with broader diversity in several classes, contrasting this method against the known limitations of GANs.

C. *Vector Quantized Diffusion Model for Text-to-Image Synthesis*

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, Baining Guo

The VQ-Diffusion model is a method for text-to-image generation. It combines the vector quantized variational autoencoder (VQ-VAE) with a conditional variant of the Denoising Diffusion Probabilistic Model (DDPM). This model eliminates the unidirectional bias and avoids the accumulation of prediction errors. It incorporates a mask-and-replace diffusion strategy, where image tokens are gradually denoised based on the input text. The VQ-Diffusion model provides global context for each token prediction and significantly improves the quality of synthesized images compared to conventional autoregressive models and previous GAN-based methods. It also allows for efficient image generation by reparameterization, achieving a better trade-off between quality and speed. To tackle the unidirectional bias problem, the model incorporates a mask-and-replace diffusion strategy. This strategy involves stochastically masking some image tokens, allowing the network to explicitly know the corrupted locations during the reverse estimation.

The VQ-Diffusion model offers several advantages over conventional autoregressive models and GAN-based methods in text-to-image generation:

1) Elimination of unidirectional bias.

2) Avoidance of error accumulation.

3) Improved image quality.

4) Efficient image generation.

Overall, the VQ-Diffusion model addresses the limitations of conventional autoregressive models and GAN-based methods, offering improved image quality, elimination of bias, avoidance of error accumulation, and efficient image generation.

D. *CogView*

Ming Ding , Zhuoyi Yang , Wenyi Hong , Wendi Zheng

CogView addresses the longstanding challenge of text-to-image generation in the general domain by introducing a potent generative model combined with cross-modal understanding. The proposed solution, CogView, is a Transformer with 4 billion parameters and employs a VQ-VAE tokenizer, aiming to push the boundaries of this problem.

In the realm of pretext tasks, text-to-image generation stands out by challenging the model to disentangle intricate elements such as shape, color, and gesture from pixel-level information. It requires the model not only to comprehend input text but also to align objects and features with their corresponding words and synonyms. Furthermore, the model must learn complex distributions to skillfully generate overlapping and composite representations of various objects and features. This process, akin to painting, transcends basic visual functions, demanding a higher-level cognitive ability.

In a recent development, DALL-E independently introduced a similar concept, preceding the release of CogView. In comparison to DALL-E, CogView advances the state of the art on several crucial fronts. CogView significantly surpasses DALL-E and previous GAN-based approaches by a considerable margin, as evidenced by the Fréchet Inception Distance (FID) on blurred MS COCO. Notably, CogView stands as the pioneering open-source large text-to-image transformer

CogView's backbone consists of a unidirectional Transformer with an impressive architecture. The Transformer comprises 48 layers, a hidden size of 2560, 40 attention heads, and a grand total of 4 billion parameters. To indicate the boundaries of text and image within each sequence, four separator tokens are incorporated. Additionally, all sequences are either clipped or padded to a uniform length of 1088.

E. Shifted Diffusion for Text-to-image Generation

Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, Jinhui Xu

The main focus of the paper "Shifted Diffusion for Text-to-image Generation" is to propose a novel diffusion model called Corgi that improves the text-to-image generation process. The paper explores techniques to enhance the diffusion process itself, making it more suitable and effective for generating high-quality image embeddings from text. The authors aim to bridge the gap between image and text modalities and train a better generative model. The Corgi model is designed to be flexible and can be applied in various settings, including supervised, semi-supervised, and language-free text-to-image generation.

The key components of the proposed framework for text-to-image generation are as follows:

1. Pre-trained Image Encoder: This component maps images to their corresponding embeddings. It is responsible for encoding the visual information of the images.

2. Decoder: The decoder generates images from the embeddings. It takes the image embeddings as input and produces the corresponding images.

3. Prior Model: The prior model generates image embeddings from text captions. It takes the text captions as input and generates the corresponding image embeddings.

These three components work together to enable the generation of images from text. The shifted diffusion model improves image embedding generation by incorporating prior knowledge from the pre-trained CLIP model into the diffusion process. Unlike the baseline diffusion model, which starts with random Gaussian noise, the shifted diffusion model initializes the diffusion process with an image embedding inside the effective output space of the CLIP image encoder. This initialization is closer to the target embedding, as revealed by the concept of modality gap in CLIP.

III. DATASET

The task of generating images from text descriptions requires datasets which consist of images with corresponding text descriptions which describe them.The goal is to map these text descriptions to the corresponding visual representations , learn from them and generate accurate images based on the descriptions provided. The dataset should contain images which are broad and diverse so that the model can generate robust representations.

The model is trained on the Flickr30k dataset which consists of over 31000 images and each image is associated with 5 different textual descriptions which provide accurate and diverse linguistic descriptions. This dataset is famous for its use in the field of natural language processing , caption generation and text to image generation.

Since each image is provided with five different descriptions , it helps to provide a different perspective to describe the images.

The data fields of the Flickr30k dataset consist of the following:

1) Images : Tensor consisting of the images.
2) Texts : Tensor which represents texts associated with images.
3) Comment_Nos : Tensor which represents the number of comments.

LAION-400M is a colossal dataset pivotal in the intersection of computer vision and natural language processing. It's a compendium of around 400 million image-text pairs, harvested from the vast expanses of the internet. This dataset serves as a foundational bedrock for training advanced AI models capable of understanding and generating visual content from textual descriptions.

The CLIP Model has been pretrained on this dataset for better results.It helps it to amplify the results. Training CLIP on this dataset helps the model learn a more comprehensive and nuanced understanding of visual content, enabling it to recognize and interpret a vast array of images more effectively.

Training on a diverse and large dataset can also help in mitigating biases. The variety in LAION-400M ensures that the model is exposed to a wide range of human and cultural perspectives, reducing the risk of overfitting to a narrow set of data characteristics.

IV. MODEL

This architecture's main concept is to take advantage of CLIP's simultaneous text and image understanding capabilities. Text is encoded into a suitable representation, and an encoder, generator (U-Net diffusion model), and decoder are used in tandem to create an image that corresponds with the text description. To make sure the generated image matches the textual input, the decoder helps refine it..

CLIP

CLIP is a model developed by OpenAI whose main goal is to derive a relationship between image and text. In our Model we are using CLIP to take in the input text description and return a fixed size vector representation.

CLIP Embeddings

CLIP embedding consists of two components:

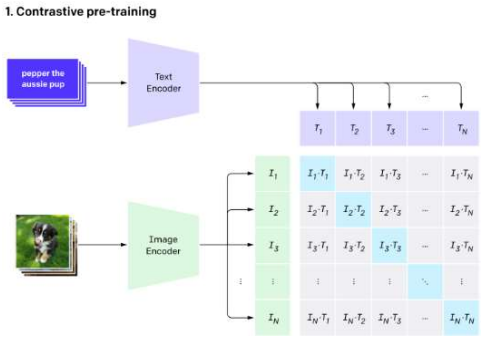
1) Token embeddings : Convert token IDs into continuous vectors.
2) Position embeddings : Encodes positional information for each token in the input sequence.

CLIP Layer

This includes self-attention, layer normalization , feedforward sub-layers.The self-attention mechanism captures contextual information within the text embeddings. The feedforward layers transform the input embeddings, adding non-linearity to the model's representations.

CLIP Main

This combines both the CLIP Layer and the CLIP embeddings. Uses CLIP Embeddings for tokenization and embedding text input. It also consists of multiple CLIP layers creating a deep network(We have used 12 layers).



Encoder

This is used to define a VAE (Variational Autoencoder) encoder architecture that takes an image as input and produces a latent representation. This encoder architecture is designed to work in conjunction with CLIP (Contrastive Language-Image Pretraining) and is used to encode images into a format that can be compared with text embeddings produced by CLIP.

The VAE encoder processes an input image with three color channels (RGB) and gradually reduces its spatial dimensions while increasing the number of channels in its feature maps. It follows a series of convolutional and residual blocks to perform this transformation.

The encoder performs downsampling using convolutional layers with a stride of 2x2. This reduces the spatial dimensions of the feature maps while doubling the number of channels. This is done twice in the network.

After the downsampling steps, the encoder applies an attention block to capture global context and dependencies within the image.

The final output of the VAE encoder is the latent representation of the input image.

UNET

The Unet iteratively denoises the latent image representations while being conditioned on the text embeddings.

The output of the U-Net, being the noise residual, is used to compute a denoised latent image representation via a scheduler algorithm.

The denoising process is repeated many times to step-by-step retrieve better latent image representations. Once complete We will be able to see the output image.

Forward Process

In the “Forward Diffusion” process, we slowly and iteratively add noise(corruption) to the images in our training set.

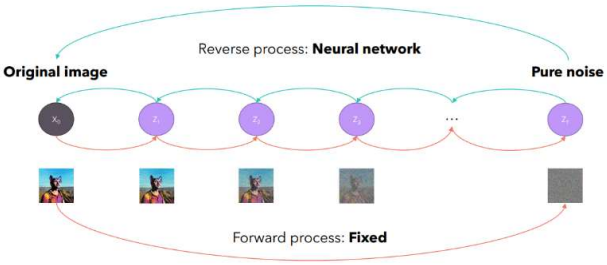
At the end of the forward process, the images become entirely unrecognizable. The complex data distribution is wholly transformed into a (chosen) simple distribution. Each image gets mapped to a space outside the data subspace.

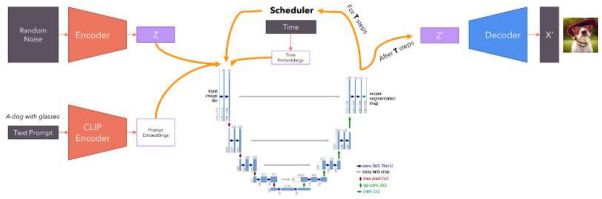
Reverse Process

The reverse process starts where the forward process ends.

We slowly and iteratively try to reverse the corruption performed on images in the forward process.

This approach for training and generating new samples is much more stable than GANs and better than previous approaches like variational autoencoders (VAE) and normalizing flows.





V. RESULTS

The following Model has been trained on these hardware specifications:

CPU : AMD Ryzen Threadripper PRO 3945WX 12-Cores

RAM: 64 GB DDR6 RAM

GPU: NVIDIA Quadro RTX 4000 GPU

1) Text Prompt : A man with a hat who looks like a clown , highly detailed

   Image Generated:

2) Text Prompt : Imagine a futuristic cityscape at dusk. The skyline is dotted with towering skyscrapers

   Image Generated:

3) Text Prompt: Generate an image of a serene beach sunset with palm trees and gentle waves

   Image Generated:

4) Text Prompt: Illustrate a cozy, book-filled library with warm lighting and comfortably seated boys.

   Image Generated:

5) Text Prompt: Mythical forest in the heart of a magical realm at twilight.

   Image Generated:

6) Text Prompt : Generate an image of an ancient castle on a cliff, surrounded by mist and overlooking a landscape.

   Image Generated:

7) Text Prompt : Abandoned buildings highly detailed

   Image Generated:

As we can see, the generated image is able to accurately capture the details from the text input. The generated image uses multiple attention layers to capture these details.

BENCHMARK RESULTS

	DALL-E†	CogView†	Lafite†	LDM(Ours)
FID ↓	27.50	27.10	26.94	23.35
IS ↑	17.90	18.20	26.02	19.93±0.35

Higher Inception Score suggests that the given model is able to generate images which are diverse and have high quality features. As seen in the table above , the model is able to outperform DALL-E and Cog-View in terms of Inception Score and is not that far behind as compared to Lafite. Lower FID Score is desired as it suggests that the generated images are closer to the real data distribution . The table suggests that the model has the least FID Score about all the above models.

According to the benchmark results the model has found to be able to generate diverse , accurate images according to the given text descriptions and the model is performing satisfactorily with a low FID and high Inception Score.

VI. CONCLUSION AND FUTURE WORK

The model and code presented here offer a remarkable convergence of text and image generation technologies. Leveraging the power of CLIP and diffusion models, this solution demonstrates the potential for creative AI-driven image synthesis based on textual prompts and input images.

The results of the model have been found to be satisfactory and have much scope of improvement in the field of generating images from text descriptions. The generated images have been quite clear and accurate and do a good job of representing the actual text input.

Future Work includes

1) Speed and Efficiency: Optimization techniques, parallel processing, and model compression should be investigated to enhance speed and efficiency, making the model more accessible.
2) Diverse Outputs: Strategies to encourage the generation of diverse images for similar prompts can be pursued, ensuring that users receive a broader range of creative options.
3) Fine-Tuning: Fine-tuning the model on domain-specific datasets can elevate its ability to generate contextually relevant and high-quality images.
4) User-Friendly Interface: Building an intuitive user interface can democratize the usage of this model, enabling non-technical users to harness its creative potential.

REFERENCES

[1] Aaron van den Oord , Oriol Vinyals and Koray Kavukcuoglu , “Neural Discrete Representation Learning”, 2017

[2] Ali Razavi, Aaron van den Oord and Oriol Vinyals , “Generating Diverse High-Fidelity Images with VQ-VAE-2”, 2019

[3] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang and Jie Tang, “CogView : Mastering Text-to-image Generation via Transformers, 2021

[4] Hyungyung Lee, Sungjin Park, Joonseok Lee, Edward Choi, “Unconditional Image-Text Pair Generation with Multimodal Cross Quantizer” , 2022

[5] Taehoon Kim , Gwangmo Song, Sihaeng Lee , Sangyun Kim , Yewon Seo , Soonyoung Lee ,Seung Hwan Kim ,Honglak Lee and Kyunghoon Bae from LG AI Research, “L-Verse : Bidirectional Generation Between Image and Text”, 2020

[6] Hiroshi Sasaki , Chris G , Willcocks , Toby P and Breckon from Department of CS from Durham University, Durham, UK , “UNIT-DDPM: Unpaired Image Translation with Denoising Diffusion Probabilistic Models “ , 2021

[7] Andrea Frome , Greg S , Jonathon Shlens , Samy Bengio , Jeffrey Dean, Marc’Aurelio Ranzato and Tomas Mikolov , “DeViSe : A Deep Visual-Semantic Embedding Model” , 2021

[8] Hongchen Tan, Xiuping Liu, Baocai Yin and Xin Li , “DR-GAN: Distribution Regularization for Text-to-Image Generation” , 2022

[9] Jianmin Bao , Fang Wen , Lu Yuan , Baining Guo , Shuyang Gu , Dong Chen , Bo Zhang and Dongdong Chen , “Vector Quantized Diffusion Model for Text-to-Image Synthesis”

[10] Sanghyuck Na , Mirae Do , Kyeonah Yu and Juntae Kim , “Realistic Image Generation from Text by Using BERT-Based Embedding” , 2022

[11] Jianmin Bao , Fang Wen, Lu Yuan , Baining Guo , Shuyang Gu, Dong Chen , Bo Zhang and Dongdong Chen , “ Vector Quantized - Diffusion model”

[12] Wilson Yan , Yunzhi Zhang , Pieter Abbeel and Aravind Srinivas , “VideoGPT: Video Generation using VQ-VAE and Transformers” , 2021

[13] Han Zhang , Tao Xu , Hongsheng Li , Shaoting Zhang , Xiaogang Wang , Xiaolei Huang and Dimitris Metaxas1 , “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks”

[14] Han Zhang , Tao Xu , Hongsheng Li , Shaoting Zhang , Xiaogang Wang and Xiaolei Huang, “StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks” , 2019

[15] Dmitrii Torbunov, Yi Huang, Haiwang Yu, Jin Huang, Shinjae Yoo, Meifeng Lin, Brett Viren, Yihui Ren Brookhaven National Laboratory, Upton, NY, USA; “UVCGAN: UNet Vision Transformer cycle-consistent GAN for unpaired image-to-image translation”; 2023

[16] Sheng Shen, Liunian Harold Li, Hao Tan , Mohit Bansal , Anna Rohrbach , Kai-Wei Chang , Zhewei Yao and Kurt Keutzer ; “How Much Can CLIP Benefit Vision-and-Language Tasks?” ; 13th July 2021

[17] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, Sungroh Yoon ; “ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models”

[18] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, Zili Yi ; “CLIP-GEN: Language-Free Training of a Text-to-Image Generator with CLIP” ; 1st March 2022
[19] Yufan Zhou , Bingchen Liu , Yizhe Zhu , Xiao Yang , Changyou Chen , Jinhui Xu ; “Shifted Diffusion for Text-to-image Generation”
[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala; “Adding Conditional Control to Text-to-Image Diffusion Models” ;