iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** X **Month of publication:** October 2023

DOI: https://doi.org/10.22214/ijraset.2023.56024

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue X Oct 2023- Available at www.ijraset.com

Text To Image Generation By Using Stable Diffusion Model With Variational Autoencoder Decoder

Usharani Budige[1], Srikar Goud Konda[2]
Department of Information Technology, BVRIT HYDERABAD College of Engineering for Women

Abstract: ***Imagen is a text-to-image diffusion model with a profound comprehension of language and an unmatched level of photorealism. Imagen relies on the potency of diffusion models for creating high-fidelity images and draws on the strength of massive transformer language models for comprehending text. Our most important finding is that general large language models, like T5, pretrained on text-only corpora, are surprisingly effective at encoding text for image synthesis: expanding the language model in Imagen improves sample fidelity and image to text alignment much more than expanding the image diffusion model.***

Keywords: Latent Diffusion Model, Imagen, U-net, Autoencoder.

I. INTRODUCTION

Across the fields of most amazing current advances in computer vision is image synthesis, which is also one of the areas with the highest computing needs. Scaling up likelihood-based models, which may contain various parameters in autoregressive (AR) transformers, currently dominates the synthesis of complex, natural landscapes at particularly high resolutions. The promising outcomes of GANs, in comparison, have been shown to be largely restricted to data with comparable low levels of variability since their adversarial learning process is not well suited to modeling complicated, multi-modal distributions. Multimodal learning has become progressively more popular in the past few years, with image-text contrastive learning and text to image synthesis at the forefront. These models' innovative image generating and editing tools have revolutionized the scientific community and attracted considerable public interest.

In order to advance this field of study, a text-to-image stable diffusion model was introduced. For text-to-image synthesis, it combines the strength of transformer language models (LMs) with high-fidelity diffusion models to achieve a level of photorealism and linguistic comprehension that is previously unheard of. The key discovery of Imagen is that, unlike earlier research that simply employs image-text data for model training, text embeddings from large LMs that were practice with text alone are remarkably effective for text-to-image synthesis. To translate input text into a series of embeddings, Imagen uses a frozen T5-XXL encoder, a 64x64 image diffusion model, and two Super Resolution diffusion models. For diverse text inputs, choose Imagen samples of 1024 x 1024. All diffusion models employ classifier-free guiding and are conditional on the text embedding sequence. Imagen relies on unique sampling techniques to produce imagery with more fidelity and better alignment of the image to the text than was previously feasible, enabling the use of large guide weights without suffering from the sample quality degradation shown in recent work. These are mostly made to turn words into beautiful graphics.

II. RELATED WORK

In image generation, diffusion models have achieved widespread success [1, 3, 4, 2], exceeding GANs in fidelity and variety while avoiding concerns with training instability and mode collapse [6, 2].

DALL-E 2 [5] uses a diffusion prior on CLIP text latents and cascaded diffusion models to generate high resolution 1024X1024 images; we believe Imagen is much simpler as Imagen does not need to learn a latent prior, yet achieves better results in both MS-COCO FID and hum analysis.

Autoregressive models [7], GANs [9, 8], VQ-VAE Transformer-based methods [10], and diffusion models have considered.

We employ huge pretrained frozen language models instead of the cascaded diffusion models that GLIDE [11] uses for text-to-image, which we found to be more effective for image fidelity and picture-text alignment.

©IJRASET: All Rights are Reserved

Although we expand to considerably larger text encoders and show their efficacy, XMC-GAN [8] also uses BERT as a text encoder. Additionally common in the literature [13], cascaded models have been successfully applied in diffusion models to produce high resolution images [2, 12].

.

III. PROPOSED METHOD

A. Model Training

We point out that even though diffusion models allow for the undersampling of the corresponding loss terms to overlook perceptually unimportant information, they nevertheless demand expensive function evaluations in pixel space, which places a significant demand on computation time and energy resources. By doing so, the computational requirements of training diffusion models for high-resolution picture synthesis will be reduced. By explicitly separating the compressive from the generative learning phases, we suggest avoiding this problem (see Fig. 3.1). To do this, we employ an autoencoding method that learns a space that is perceptually similar to the picture space but has a significantly reduced computing complexity. Due to the convolutional architecture of our suggested LDMs, which scales more easily to larger dimensions latent spaces, such compromises are avoided. As a result, we have the freedom to choose the amount of pressure that, for a phase, best mediates between learning the main regions of strength, without overtaxing the generative dissemination model with perceptual pressure, while still maintaining accurate constancy reconstructions. Even if there are ways to become familiar with an encoding model couple with an earlier result based either jointly or independently.

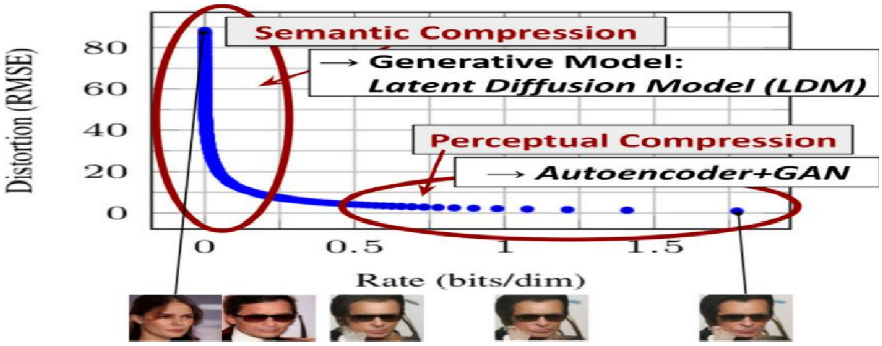


Fig 3.2.1 Examples of semantic and perceptual compression

B. Pretrained text encoders

In order to know the complexity and compositionality of natural language text inputs, for text to image models require robust semantic text encoders. Current text to image models come prepackaged with text encoders that have been pretrained on paired image from text data or trained from scratch. According to the image from text training objectives, these text encoders may encode visually semantic and respective representations that are particularly pertinent for the task of converting text into images. Another way to encrypt text for text to image generation is a large language model. New developments in huge language. Advances in textual comprehension and generative capacities have been made because to models. Language models are exposed to a very rich and diverse distribution of text since they are trained on text-only corpora, which are substantially larger than coupled image-text data. Additionally, these models are typically a lot bigger than text encoders in existing image-text models. Therefore, investigating both families of text encoders for the text to image job makes sense.

C. Imagen explores pretrained text encoders

We set text encoder weights to zero. Freezing provides a number of benefits, including the ability to compute embeddings offline, which requires little processing power or memory during the training of the text-to-image model.

D. Diffusion Models

The process of diffusion models are probabilistic models created to learn a data distribution p(x) by gradually denoising a normally distributed variable, which equates to learning the opposite process of a fixed T-length Markov Chain.

The most effective models for picture synthesis rely on a reweighted version of the variational lower bound on p(x), which mimics denoising score-matching.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue X Oct 2023- Available at www.ijraset.com

E. U-net

Both the encoder and the decoder components of the U-Net are made up of ResNet blocks. A higher resolution image representation that is purportedly less noisy is converted by the encoder into a lower resolution image representation, which is then converted by the decoder back to the original. More specifically, the anticipated denoised picture representation can be computed using the noise residual predicted by the U-Net output.

F. Imagen

To develop Imagen, a text encoder turns text into a number of embeddings. A number of conditional diffusion models then take these embeddings and produce images with ever greater resolutions(Fig. 3.2). We go into great detail about each of these elements in the subsections that follow.

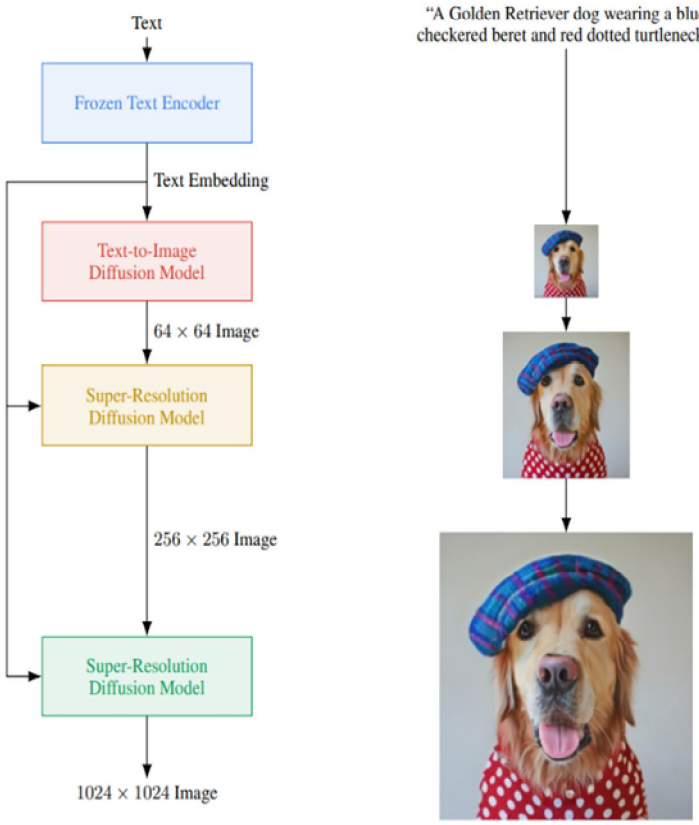


Fig 3.2: Visualization of Imagen.

IV. RESULT

For the 64x64 text-to-image synthesis, we train 2B parameter models with 600M and 400M parameter models for 64x64, 256x256, and 1024x1024 for super resolution, respectively. For all models, we employ a batch size of 2048 and 2.5M training steps. We use 128 TPU-v4 chips for our two super-resolution versions and 256 TPU-v4 chips for our 64 x 64 basic model.

Being physically fit is not a problem in our opinion, and additional training may enhance performance in general. We choose Adafactor for our base 64 x 64 model because initial comparisons with Adam showed similar performance with a noticeably reduced memory footprint for Adafactor. We choose Adam for super resolution models since, in our initial ablations, we found Adafactor to be deleterious to model quality. By zeroing out the text embeddings with a 10% probability for each of the three models, joint and train unconditionally for classifier-free guiding. For training, we merge 460M image-text pairs from our own datasets with 400M image to text pairs from the publically available Laion dataset.

©IJRASET: All Rights are Reserved

TABLE I
calculating the Model performances

		Zero -shot
Model	**FID-30K**	**FID-30**
AttnGAN	35.49	
DM-GAN	32.64	
DF-GAN	21.42	
DM-GAN+CL	20.79	
XMC-GAN	9.33	
LAFITE	8.12	
Make-A-Scene	7.55	
DALL-E		17.89
LAFITE		26.94
GLIDE		12.24
DALL-E 2		10.39
Imagen(Total work)		**7.27**

A. Input Output results

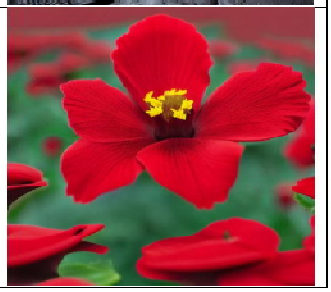
TEXT(input)	Output
Bike riding on moon	
Flower is red in color,with petals are red in color and bunched together	

Fig5.1 Input output Results

V. CONCLUSION

Imagen shows how to use text encoders that are massive, frozen pretrained language models for diffusion model-based text-to-image generation. Further research into the usage of increasingly larger language models as text encoders is motivated by our discovery that increasing the size of these language models has a noticeably greater impact on overall performance than increasing the size of the U-Net. We also introduce dynamic thresholding and underline the significance of classifier-free guiding in Imagen, which permits the use of much higher guidance weights than in prior work. Imagen creates 1024 X 1024 samples with exceptional photorealism and text alignment using these innovative components with Imagen, our main goal is to improve the study of generative techniques while employing text-to-image synthesis as a test case. We acknowledge the possible downstream uses of this study are diverse and may have a significant impact on society, even though end-user applications of generative approaches are still mainly outside of reach. Generative models have the ability to greatly enhance, extend, and complement human creativity. Especially text-to-image providing models have a chance to expand the possibilities of picture editing and result in the creation of fresh resources for professionals in the creative industry. However, generative methods raise a lot of questions about prejudice and exclusion from society because they can be used for bad things like harassing and spreading false information. These factors influenced our choice not to make code or the public demo available. We will investigate an approach for safe externalizing that strikes a compromise between the advantages associated with outside audits and the disadvantages of unrestrained open access in the future research. A majority of text-to-image approaches in use today ignore the difference between foreground and background in order to create images in a comprehensive manner, which causes items in images to be easily disturbed by the backdrop. Furthermore, they typically fail to recognize the mutual dependence of several generative model types, Using latent diffusion models is a quick and easy way to increase the training and despite the fact that LDMs require significantly less computing power than pixelbased methods.

The reconstructed capabilities of our models can be a bottleneck for applications that demand fine-grained precision in pixel space, despite the fact that there is relatively little picture quality loss in our models. Additionally, this is one area in which our super-resolution models are already a little bit limited. As evidenced by the results, background elements can be enhanced in order to produce better results consistent with the text and human picture quality. A quick and easy technique to increase the training and sampling efficiency of denoising diffusion models without affecting their quality is to use latent diffusion models.

VI. ACKNOWLEDGMENT

We are grateful to Ben Poole for reading our text, participating in early conversations, and offering numerous insightful comments and recommendations all throughout the project. We would especially want to thank Sarah Laszlo, Austin Tarango, and Kathy Meier-Hellstern for their assistance in assisting us in integrating significant ethical AI principles into our project. Elizabeth Adkison, Zoubin Ghahramani, Jeff Dean, Yonghui Wu, and Eli Collins have all provided us with insightful comments and assistance. The Imagen watermark was created by Tom Small, for which we are grateful. For the first talks and comments, we are grateful to Jason Baldridge, Han Zhang, and Kevin Murphy.

Victor Gomes and Erica Moreira's constant and indispensable assistance with TPU resource allocation is gratefully acknowledged. Additionally, we would like to thank Shekoofeh Azizi, Harris Chan, Chris A. Lee, and Nick Ma for giving up a sizable portion of their time to test DrawBench.

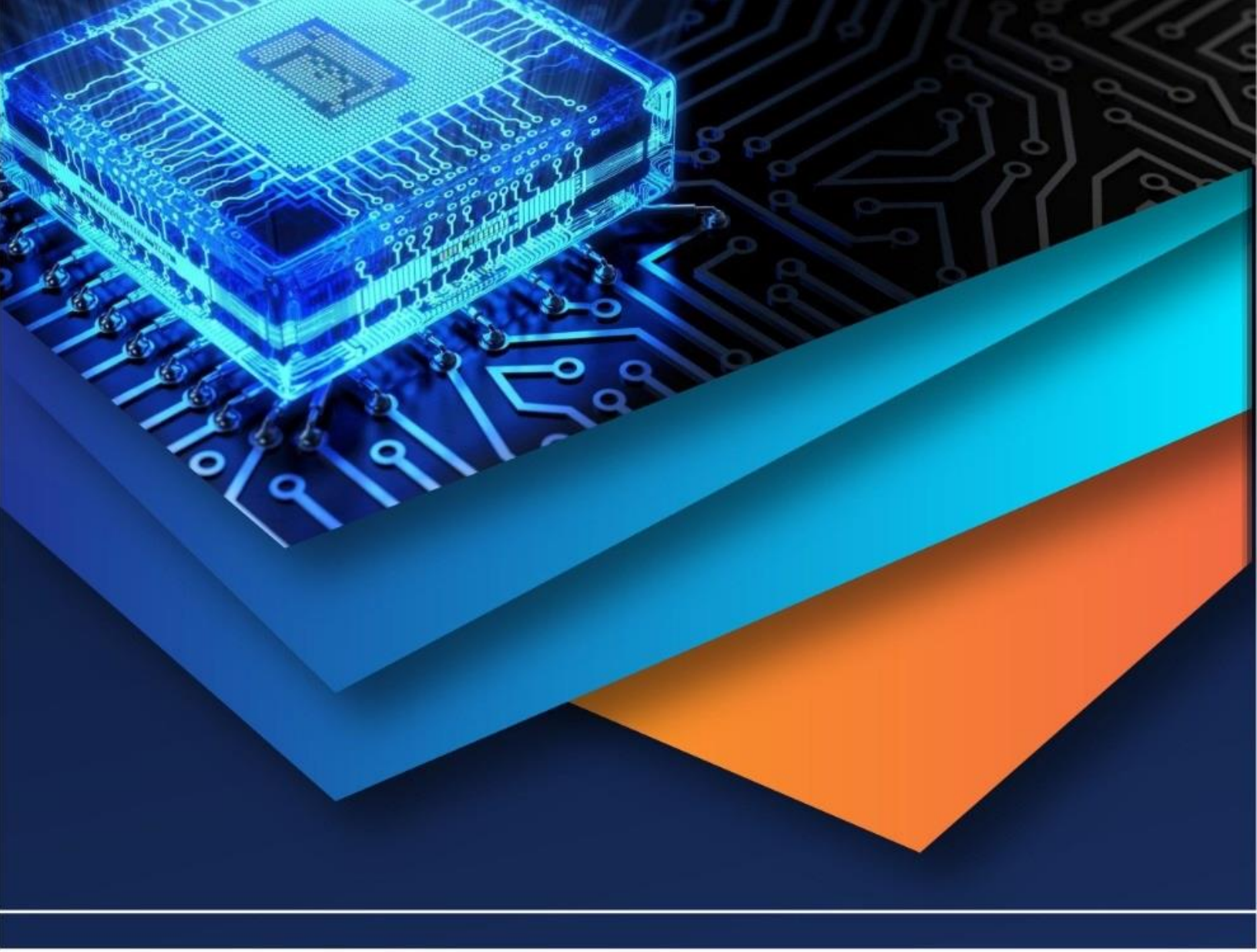
REFERENCES

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. NeurIPS, 2020.
[2] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In NeurIPS, 2022.
[3] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. arXiv preprint arXiv:2102.09672, 2021.
[4] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. arXiv preprint arXiv:2104.07636, 2021.
[5] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. In arXiv, 2022.
[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096, 2018.
[7] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating Images from Captions with Attention. In ICLR, 2016.
[8] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-Modal Contrastive Learning for Text-to-Image Generation. In CVPR, 2021.
[9] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In CVPR, 2018.
[10] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In ICML, 2021.

©IJRASET: All Rights are Reserved

[11] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Bob McGrew Pamela Mishkin, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2021.
[12] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. JMLR, 2022.
[13] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In NIPS, 2015.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)