# TEXT TO IMAGE GENERATION USING STABLE DIFFUSION

**Divyanshu Mataghare[1], Shailendra S. Aote[2], Ramchand Hablani[3]**

**Article History:** **Received:** 19.03.2023 **Revised:** 03.05.2023 **Accepted:** 17.06.2023

**Abstract**

Diffusion models (DMs) provide cutting-edge synthesis outcomes on image data and beyond by breaking down the picture generation process into a sequential application of denoising autoencoders. Furthermore, their design enables a guiding system to regulate the picture generating process without retraining. Nevertheless, because these models frequently work in pixel space, optimization of strong Because to sequential assessments, DMs sometimes need hundreds of GPU days, and inference is costly. We use them in the latent space of potent pretrained autoencoders to empower DM preparing on obliged processing assets while keeping up with their excellence and adaptability. Contrary to earlier research, using such a representation to train diffusion models enables for the first time to achieve a nearly ideal balance between the preservation of detail and complexity reduction, significantly enhancing visual fidelity. By including cross-attention layers into the model architecture, we convert diffusion models into powerful and flexible generators for common conditioning inputs like text or bounding boxes and enable high-resolution synthesis in a convolutional manner.For picture inpainting and class-restrictive picture blend, inert dissemination models (LDMs) accomplish new cutting edge scores. incredibly serious execution on a scope of undertakings, including as text-to-picture blend, unrestricted picture creation, and super-goal, while requiring significantly less handling power than pixel-based DMs.

**Keywords:** Image Generation, Deep Learning, Stable Diffusion

[1,2,3]Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Email: shailendra.aote@gmail.com

**DOI: 10.31838/ecb/2023.12.s3.496**

## 1. Introduction

Photo editing, computer-aided design, and other fields can all benefit greatly from the ability to create photo-realistic visuals from text. Current research on the synthesis of real-world photographs using Generative Adversarial Networks (GAN) has demonstrated encouraging results. Convolutional Deep GAN (DC-GAN). A GAN model called DC-GAN [6] uses deep CNN for each generator and discriminator model. This structure essentially uses CNNs to produce images from noisy data that fall into a certain distribution using the Generator-Discriminator framework. Nevertheless, the networks are only trained for one step each, alternately. A conditional generative model may be created by simply converting GANs, as described in the original GAN paper[1]. The generator and discriminator are both added to in either of their levels to produce data that is conditional on a condition vector c. The networks will develop the ability to modify their parameters in response to these new inputs. A probabilistic graphical model may also be used to view conditional GANs. With standard GANs, the noise Z affects the observable X. In the special case of text-to-picture synthesis, the states e of C are vectors encoding a text description. With conditional GANs, X is influenced by both Z and C. The topic of text to picture synthesis was initially addressed by Reed et al. [2] with promising results.Developing an outwardly discriminative portrayal for the message depictions and using this Portrayal to deliver practical pictures are the two key subproblems that make up the overall challenge.

A StackGAN gets its name from the fact that it consists of two GANs that are stacked together to create a network that can produce high-resolution pictures. Stage I and Stage II are its two phases. Whereas the Stage-II network uses the picture created by the Stage-I network to create a high-resolution image that is dependent on a text embedding, the Stage-I network creates low-resolution images with basic colours and crude drawings. In essence, the second network fixes flaws and adds attractive features to produce a higher-resolution image that is more lifelike. Microsoft Research created AttnGAN [3] in association with other academic institutions. As the name implies, StackGAN-v2 is an enhanced version of StackGAN that employs several generators and discriminators in a tree-like structure. The design of AttnGAN is comparable to that of StackGAN-v2 [4], but it also includes an attention model on top of it. The network can concentrate on a single word from a sentence or a particular area of an image at a time using the attention model, which simulates the human attention process.GAN Wasserstein (WGAN). The WGAN [5] is an extension of the generative adversarial network 4 family that increases model stability and specifies a loss function that accounts for the measure of the difference between the probability distribution of actual and fake pictures. Instead of displaying the likelihood that something is real or false, the Critic has been modified to produce a realness/fakeness score. Examination of currently prepared dissemination models in pixel space is the most vital phase in the flight to dormant space strategy.Learning may be separated into two stages, broadly speaking, as with any likelihood-based model: The first stage is perceptual compression, which eliminates high-frequency information while learning just a small amount of semantic diversity. The real generative model picks up on the semantic and conceptual makeup of the material in the second step (semantic compression). In order to train diffusion models for high-resolution picture synthesis, we must first choose a perceptually comparable but computationally more 6 appropriate space.

**Related work**

Generative models for synthesis of images Generative modelling faces unique difficulties due to the high dimensionality of pictures. Networks of Generative Adversaries(GAN) [1] provide effective examining of high goal pictures with adequate perceptual quality [7], however, they are trying to tune [8] and have trouble capturing the complete data distribution. In contrast, likelihood-based techniques priorities accurate density prediction, making optimisation more compliant. High resolution pictures may be synthesised well using variational autoencoders (VAE) [9] and flow-based models [10], but sample quality is not on par with GANs. A sequential sampling procedure and computationally costly designs limit the resolution of the pictures that autoregressive models (ARM) [11] can produce, despite their good performance in density estimation. Maximum-likelihood training uses a disproportionate amount of capacity to model the scarcely perceptible, high-frequency features that are present in pixel-based representations of pictures, leading to lengthy training timeframes.Many two-stage techniques model a compressed latent image space with ARMs rather than raw pixels in order to scale to higher resolutions.Recent advancements in sample quality and density estimation [13] have been made by Diffusion Probabilistic Models (DM) [16]. When these models' neurological underpinnings are implemented as UNets, they naturally suit the inductive biases of image-like data, which gives rise to their generative capacity. When a re-weighted goal is used for training, the best synthesis quality is often attained. In this present circumstance, the dissemination Model is comparable to a lossy blower and considers the

compromise of pressure proficiency for picture quality.Nevertheless, the disadvantage of evaluating and improving these models in pixel space is low inference speed and very large training costs. Although improved sampling techniques and hierarchical sampling can help to some extent with the former,at approaches.Preparing on high-goal picture information generally expects to ascertain costly angles. We address the two downsides with our proposed LDMs, which work on a compacted inactive space of lower dimensionality.Image Synthesis in Two Stages Several studies [12] have focused on using a two step strategy to combine the benefits of various techniques into more effective and performant models in order to reduce the drawbacks of individual generative approaches. Auto-regressive models are used by VQ-VAEs to develop an expressive prior over a discretized latent space. By studying a combined distribution across discretized picture and text representations,extend this method to text-to-image creation.In contrast to VQ-VAEs, VQGANs [12] scale auto-regressive transformers to bigger pictures using a first stage with an adversarial and perceptual goal.

The overall performance of such techniques is, nonetheless, obliged by the enormous pressure rates important for viable ARM preparing, which adds billions of teachable boundaries, and less pressure comes at the punishment of a high computational expense. Because to our proposed LDMs' convolutional backbone, which scales more easily to larger dimensions latent spaces, such compromises are avoided.So, we are allowed to choose the amount of pressure that best intercedes between learning major areas of strength for a phase, without giving the generative dissemination model an excess of perceptual pressure, while yet guaranteeing high constancy reconstructions.While there are strategies to become familiar with an encoding/unraveling model couple with a result-based earlier either mutually or independently.

**Proposed work**

We note that despite the fact that dissemination models permit to disregard perceptually unessential subtleties by under sampling the loss terms that correspond, They still call for expensive function evaluations in pixel space, which puts a significant strain on the resources of energy and computation time. This is to reduce the amount of computing required to train diffusion models in order to create high-resolution images. We suggest eliminating this flaw by explicitly separating the compressive from the generative learning phases(fig 1) . We use an auto-encoding approach to do this,which learns a space that is nearly identical to the image space in terms of perception but has a much lower computational complexity.

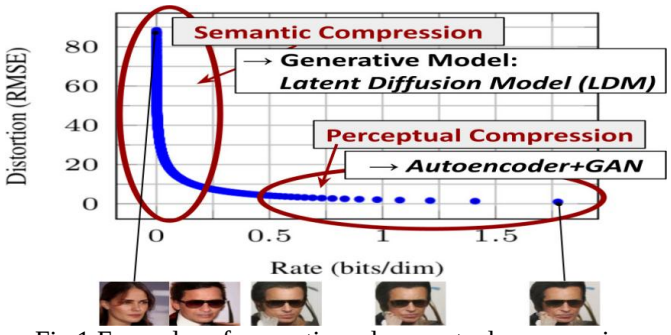


Fig 1 Examples of semantic and perceptual compression

Such a strategy has the following benefits: I By leaving the high-dimensional picture space, we are able to get DMs that employ sampling on a low-dimensional space, which is significantly more computationally efficient. (ii) We take use of the bias of induction that DMs have Its UNet architecture is inherited [14], which results in them especially useful for spatially structured data without the need for high compression levels that compromise quality while called for by other techniques. iii) As a last step, we produce general-purpose compression models, which may be applied to single-image CLIP-guided synthesis [15] as well as other downstream applications where their latent space can be used to train numerous generative models.

**Compression of perceptual images**

Our perceptual compression model, which is based on prior work , entails an autoencoder trained using a mix of an adversarial goal and a perceptual loss patch-based. This prevents bluriness from being generated by just relying on pixel-space losses, such as those caused by L2 or L1 objectives, and guarantees by ensuring local realism, the reconstructions are restricted to the picture multiple. More specifically, the encoder E converts x into a latent value given an image that is $x \in R^{H*W*3}$ in RGB space. We test two distinct types of regularisations to prevent arbitrarily high-variance latent spaces. The first form, KL-reg., employs a vector quantization layer within the decoder while applying a modest Similar to a VAE, the KL-penalty reduces the learned latent to a conventional

normal. The decoder absorbs the quantization layer in this architecture, which may be understood as a VQGAN. We can accomplish extremely accurate reconstructions while using very low compression rates since our following Our learned latent space's two-dimensional structure is made for interaction with DM z = E(x). This is different from earlier works .

**Models of Latent Diffusion**

Dissemination Models are probabilistic models made to progressively denoise a regularly dispersed variable to become familiar with an information circulation p(x), which is comparable to learning the contrary course of a decent Markov Chain of length T. The best models for picture blend utilize a reweighted variety of the variational lower limit on p(x), which is like denoising score-coordinating. One way to think of these models is as a collection of equally weighted denoising autoencoders (xt,t); t = 1....T, which have been taught to anticipate a denoised form of their input xt, where xt is a noisy form of the input x. It is possible to condense the relevant goal to with t uniformly sampled from {1;.....;T}.

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t,t)_2^2\|\right]$$

Idle Portrayal Generative Demonstrating We presently approach a successful, low-layered dormant space in which high-recurrence, imperceptible data are preoccupied away utilizing our prepared perceptual pressure models made out of E and D. This space is more appropriate for probability based generative models than the high-layered pixel space since it permits them to I focus on the pivotal, significant parts of the info and (ii) train in an extensively lower layered, computationally undeniably more productive environment. In contrast to previous work that used autoregressive, attention-based transformer models in a highly compressed, discrete latent space, we may benefit from our model's image-specific inductive biases[12]. Diffusion models, like other generative models, may theoretically represent conditional distributions of the kind p. (z|y). This opens the door to regulating using inputs like text, semantic maps [16], or other image-to-image translation tasks in the synthesis process and may be accomplished with a conditional denoising autoencoder $\epsilon_\Theta$ ($z_t$, t, y). Nevertheless, integrating the generating potential of DMs with conditionings other than class names [15] or obscured varieties of the information picture is at this point a neglected field of concentrate with regards to picture combination.

By adding the cross-attention mechanism [17] to the basic UNet backbone of DMs, which is useful for learning models of multiple input method based on attention, we make DMs become more adaptable conditional image generators.

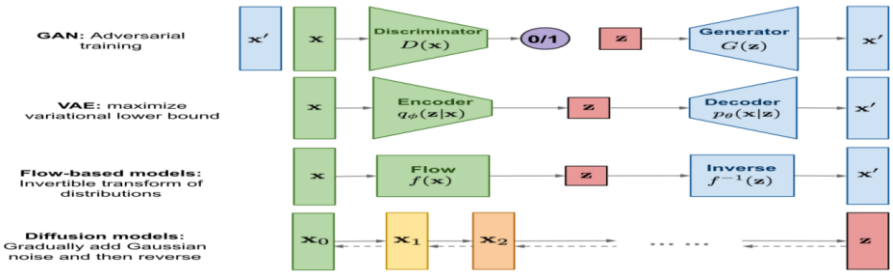


Fig 2. Overview of different types of generative models.

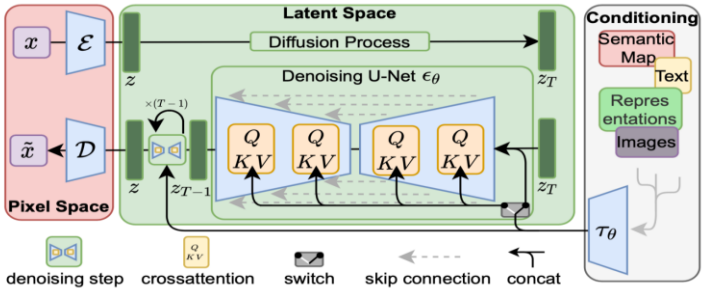


Fig. 3. The architecture of latent diffusion model.

## 2. Results

This section shows the result of our model in terms of inception score. It is found that the proposed model performs well. Fig. 5 shows the sample output.

| Ref. | Model | Inception Score(IS) |
|---|---|---|
| Reed et al. [8] | GAN-INT-CLS | 2.67±0.02 |
| Zhang et al. [16] | StackGAN | 3.21± 0.03 |
| Zhang et al. [17] | StackGAN++ | 3.25±0.02 |
| Zhang et al. [18] | HDGAN | 3.47± 0.06 |
| Cai et al. [19] | DualAttn-GAN | 4.04± 0.01 |
| **Our Proposed Method** | **LDM using stable diffusion** | **5.2± 0.05** |

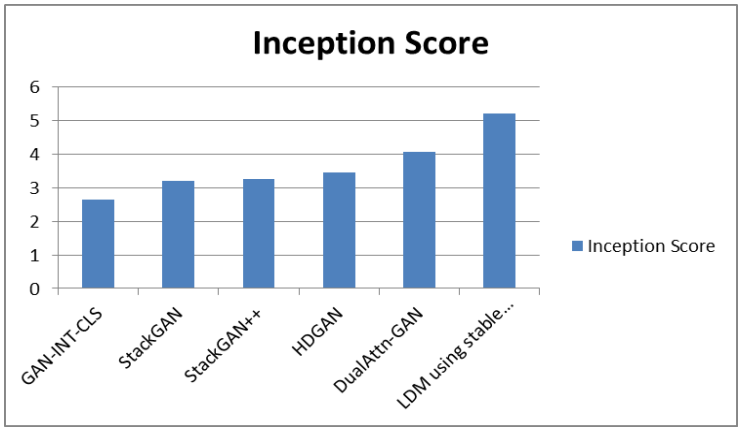


Fig 4. Bar Chart of Inception Score

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

| | |
|---|---|
| golden bird with red big beak |  |
| golden laptop on fire |  |

Fig. 5: Sample Outputs

## 3. Conclusion & Future Scope

Creating high-quality pictures from text descriptions is an interesting research topic with many practical applications. However, it is rather difficult since real-world language and visual descriptions are chaotic and highly variable. The majority of text-to-image techniques now in use seek to create images in a holistic way, ignoring the distinction between foreground and background, which leads to objects in images being readily disrupted by the backdrop. Additionally, they frequently overlook how diverse generative model types complement one another. De-noising diffusion models can have their training and sampling effectiveness improved without compromising their quality by using latent diffusion models, which are a quick and simple method. In the absence of task-specific designs based on this and our cross attention conditioning mechanism, our research may outperform current methods on a variety of conditional image synthesis tasks. Despite the fact that the sequential sampling process is still slower with LDMs than it is with GANs, despite the fact that LDMs require significantly less computing power than pixel-based methods. Despite the fact that there is very little loss of picture quality in these models, the reconstruction capabilities of our models can be a bottleneck for applications that require fine-grained precision in pixel space. We presume that this is one area where our super-resolution models are already somewhat constrained. Background things can be improved as shown in result it need to give better result according to text and image quality of human can be improved

## 4. Reference

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. He/shelling, C. Cortes, N.D. Lawrence, and K. Q. Iinberger, editors, Advances in Neural Information Processing Systems 27, pages2672{2680. Curran Associates, Inc., 2014.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text-to-image synthesis. In Proceedings of The 33rd International Conference on Machine Learning, 2016.

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. CoRR,abs/1711.10485, 2017.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N.Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. CoRR,abs/1710.10916, 2017.

Arjovsky, M., Chintala, S., & Leon, B (2017). Wasserstein GAN. arXiv:1701.07875

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H (2016). Generative adversarial text-to-image synthesis, 33rd International Conference on International Conference on Machine Learning. ICML 2016, vol. 48, pp. 1060–1069.

42_9 Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. CoRR, abs/1912.04958,2019.

Lars M. Mescheder. On the convergence properties of GAN training. CoRR, abs/1801.04406, 2018.

Diederik P. Kingma and MaxWelling. Auto-Encoding Variational Bayes. In 2nd International Conference on Learning Representations, ICLR, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017.

A¨aron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. CoRR,abs/1601.06759, 2016.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge,and Yonghui Wu. Vector-quantized image modeling with improved vqgan, 2021.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. CoRR,abs/2107.00630, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet:Convolutional networks for biomedical image segmentation. In MICCAI (3), volume 9351 of Lecture Notes in Computer Science, pages 234–241. Springer, 2015.

Kevin Frans, Lisa B. Soros, and OlafWitkowski. Clipdraw:Exploring text-to-drawing synthesis through languageimage encoders. ArXiv, abs/2106.14843, 2021.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan,and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. CoRR, abs/1503.03585, 2015.

Shailendra S. Aote, Dr M M Raghuwanshi, Dr. Latesh Malik, "A New Particle Swarm Optimizer with Cooperative Coevolution for Large Scale Optimization", Proceedings of FICTA: Springer- AISC, Vol.327, ISBN NO: 978-3-319-1933-5,14th - 15th Nov 2014, pp 781-790, DOI: https://doi.org/10.1007/978-3-319-11933-5_88.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. arXiv 2016,arXiv:1605.05396.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.;Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy 22–29 October 2017; pp. 5907–5915.

Zhang, Z.; Xie, Y.; Yang, L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6199–6208.

Cai, Y.; Wang, X.; Yu, Z.; Li, F.; Xu, P.; Li, Y.; Li, L. Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network. IEEE Access 2019, 7, 183706–183716.