iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** II **Month of publication:** February 2026

DOI: https://doi.org/10.22214/ijraset.2026.77372

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 14 Issue II Feb 2026- Available at www.ijraset.com

A Comparative Analysis of Text-to-Image Generation Models: Diffusion and Autoencoder Approaches

Sumedha Arya

Abstract: ***Text-to-image generation is a technique of creating images based on text descriptions. Recently, so many research publications has been done in this area, showing its popularity. In this work, we reviewed various autoregressive models, non-autoregressive models, GANs, energy-based models, multimodal methods and diffusion models used for text to image generation tasks. We also discuss important techniques commonly used in these models, such as autoencoders, attention mechanisms, and classifier-free guidance. For application, we performed a comparative analysis of diffusion and autoencoder models for text to image generation tasks taking Flowers-HD5 dataset. The results shows that the autoencoder achieves rapid convergence and significantly lower reconstruction loss (~0.01 range), producing sharp and faithful results. While the diffusion model, despite higher loss (~0.1–0.25), generates images with greater diversity.***

Keywords: Text-to-Image Generation, Diffusion Models, Conditional Autoencoder, Classifier-Free Guidance, Cross-Attention, Multimodal Learning.

I. INTRODUCTION

With the rapid advancement in generative AI techniques such as GANs, autoregressive (AR) models, non-autoregressive (NAR) models, and diffusion models, there is a tremendous improvement in text-to-image generation. As a result, many research papers have been published recently in this area.

Text to Image generation is conditional in nature. It depends on the type of input given as a text. Earlier models that generate images depends on captions, such as the DRAW model [28]. However, modern text to image generation research gained momentum from 2016 [60]. Nowadays, text to image generation uses foundation techniques such as Autoregressive (AR), Non-Autoregressive (NAR), GAN, and Diffusion models.

Autoregressive models generate images in steps, just like writing a sentence one word at a time. They follow the chain rule of probability, where each new part depends on all previous parts [3, 36]. Such techniques are primarily used for text generation in language models. Later it is adapted in image generation also. Transformer architecture [100] greatly improved AR models and became the backbone for GPT. For images, iGPT [9] was used that converts pixels into a long sequence and then predicts them one by one. Vision Transformer (ViT) [18] improves this idea by splitting images into patches instead of individual pixels. As a result, High quality images can be generated using AR models. However, the generation of images are slower in nature because steps cannot be fully parallelized.

Because AR models are slow, non-autoregressive models were introduced. They generate many parts of the image at the same time. This idea first came from NLP [29] and significantly speeds up inference. Models like CMLM [26] predict masked parts of an image in parallel. Later on, it refines the results over multiple iterations. Therefore, approaches based on NAR have been applied to achieve faster image generation [5, 6, 17, 23, 68]. However, sometimes the quality of image generated by NAR is poor in nature.

GANs consist of two neural networks that performs training in competition with each other. That's why it is called as adversarial in nature [12]. These two neural networks are:

- Generator (G): That creates fake images from noise and text
- Discriminator (D): That tries to distinguish between real and fake images

The first architecture based on GAN that performs text to image generation was GAN-INT-CLS [78]. GANs can produce sharp and realistic images, but their training is often long and unstable. This may cause model collapse [83].

Diffusion models are currently the most popular approach for text to image generation. They work in two stages such as forward and reverse process. In forward process, they gradually add noise to an image until it becomes pure noise. While in reverse process, they learn to remove noise step by step to recover the image, often using a U-Net [80].

This idea was introduced in DDPM [32]. However, it requires multiple steps to perform. Later, latent diffusion models (LDM) [79] came, which have made this process faster by working in a compressed latent space. This speed up the image generation, followed by DDIM [90] using fewer steps. According to the latest research, text to image generation is performed by continuous-time models such as SDE-based diffusion [84, 92], ODE-based methods [54], and Rectified Flow [54]. Distillation methods such as consistency models [58, 91] reduce sampling to very few steps. Diffusion models are widely used in text to image because of their training stability and preventing model from collapse, unlike GANs [83]. Also, text information dictates the denoising process through embeddings, often using classifier-free guidance [33].

The key strategies used in text to image generation are as follows:

1) Autoencoder compress images into smaller latent representations which are further used for reconstruction. This speed up the process of training and inferencing. Many text systems use Variational Autoencoders (VAE) [41, 42] or vector-quantized models such as VQ-VAE [99], VQ-VAE-2 [77], and VQ-GAN [21]. On the text side, words are tokenized such as BPE [87], WordPiece [15] and encoded using models such as BERT [15], T5 [74], or CLIP [73].

2) Attention mechanism further allows models to focus on relevant or important relationships between words and image regions [1, 40]. Transformers [100] use self-attention to model long-range dependencies present in the data. In text to image generation, cross-attention aligns text tokens with image patches, as used in DALL·E [76].

3) Classifier-Free Guidance is a simple and powerful technique used in diffusion-based text to generation models [33]. During training, the model learns both conditional and unconditional generation. It means with text and without text. Therefore, during inferencing, these two outputs were combined using a method in form of guidance that controls how strongly the image follows the text. Higher guidance improves image generation to align according to text but may cause visual challenges.

This research aims to explore the text to image generation techniques. By implementing and comparing a variety of models, such as diffusion and autoencoders, this study evaluates which model is better for text to image generation tasks. The paper is further divided into following sections; review of the literature, research methodology, results analysis, conclusion and future work.

II. LITERATURE REVIEW

In this section, we detailed about the various techniques used in text to image generation. It is a key area of artificial intelligence where machines create images from text descriptions. Recent advancement in deep learning models understand text well and convert them into visual content. The key techniques used in text to image generation are:

1) Autoencoders (AEs) are a core building block of many texts to image generation models. They comprise of two main parts: encoder and decoder. An encoder part compresses images into small latent vectors while a decoder part reconstructs images from those vectors. It means, meaningful image features are stored in latent space representation, which are further used for creating diverse images [41, 42]. This process is faster and more efficient.

2) Variational Autoencoders (VAEs) were early generative models. They add randomness to the data by using KL divergence and sampling techniques, instead of directly copying the inputs [41, 42]. Later, vector-quantized models such as VQ-VAE were introduced that encodes images using a discrete codebook [99, 27]. The various extensions of them are: VQ-VAE-2 for multi-level encoding [77] and VQ-GAN, for image quality improvement using perceptual loss and patch-based discriminators [21].

3) Other important image encoders include specialized models such as dVAE [96], VQ-Diffusion [30], RQ-VAE [46], Transformer-VQ [53], VQ-SEG and VQ-IMG [25], ViT-VQGAN [110], and encoders from CLIP [73].

4) For text as an input, words are first split into tokens. The methods used for it are BPE [87], WordPiece [15], SentencePiece [45], or UnigramLM [44]. These tokens are then encoded using models such as BERT [15], T5 [11, 74], or CLIP's text encoder [73].

5) Attention mechanism helps models to focus on relevant parts of text or images, even if they are far apart from each other in the sequence [1, 40]. That means, it identifies the long-range dependencies in text and images for generative models. Self-attention, was introduced in Transformer architecture, which connects all positions in a sequence directly [100].

6) In text to image generation models, attention mechanism is also very important. For example, DALL-E model uses three types of attention. These are text to text attention, image to text attention and image to image attention. This improves alignment between text and images using cross-attention [76].

7) Classifier-Free Guidance (CFG) improves image quality in diffusion models [33]. In this process, some text conditions are randomly removed and replaced with an empty token while performing training. Therefore, during inferencing time, the model mixes conditional predictions (with text) and unconditional predictions (without text). Presence of a guidance scale controls how strongly the image follows the text, pushing results closer to the conditional prediction [33].

Image Generation Paradigms

Text to image models is grouped together on the basis of how they generate images. Some models belong to more than one group.

Autoregressive (AR) Models

- AR models generate images in steps. They predict one token at a time. Early models like iGPT [9], PixelCNN [97], PixelRNN [98], and Image Transformer [67] were based on AR using small datasets. Later models were scaled up to more stronger models such as:
- DALL-E: 12B parameters, 250M image-text pairs [76]
- CogView: 4B parameters, 30M pairs, with training stability tricks [16]
- Parti: 20B parameters, high-quality images, iterative prompt refinement [111]
- M6: focuses on multi-modal pre-training [51]

However, to reduce one-directional bias, bidirectional AR models were introduced. These are:

- ImageBART [20]
- ERNIE-ViLG [113]

For multi-modal data analysis, 3D Transformers such as NÜWA handle text (1D), images (2D), and video (3D) using 3D Nearby Attention [103]. The Training Strategies used in AR models are End-to-end for ERNIE-ViLG [113] and Multi-modal pre-training for M6 [51]. ·

Non-Autoregressive (NAR) Models NAR models generate multiple parts of an image at once instead of performing in steps. Some of them are:

- MaskGIT that uses bidirectional masked prediction [6]
- Multi-stage generation appears in CogView2 [17], Muse [5], aMUSEd [68]
- Emage introduces additional improvements [23]

Their applications include super-resolution in CogView2 [17], Muse [5], and aMUSEd [68]. · Generative Adversarial Networks (GAN) GANs generate images using a generator–discriminator architecture. Both generator and discriminator are neural networks that train in a competition. Some of the models based on GANs are:

- Large-scale models: GigaGAN [39]
- One-stage GANs: DF-GAN [95]
- Improved designs: StyleGAN-T [85], GigaGAN [39], XMC-GAN [112]

Based on Prompt and Embedding, some GANs models are:

- Multi-granularity text-image matching: DAE-GAN [81]
- Pseudo-text methods: LAFITE [114]
- CLIP-based fusion: FuseDream [55], VQGAN-CLIP [13]
- Better semantic alignment: MA-GAN [109]

For training, the common strategies include:

- Segmentation-assisted training: TReCS [43]
- Modified loss functions: XMC-GAN [112]

The applications of GANs are super-resolution which is achieved in MA-GAN [109], DAE-GAN [81], GigaGAN [39].

- Diffusion Models

Diffusion models are better than both GANs and Autoencoders in terms of high-quality image generation. They add noise to the images in encoder part and gradually remove noise from decoder part to generate images. Some of the diffusion-based models are:

- Flow-based diffusion: InstaFlow [56], SD3 [19]
- Latent Diffusion Models (LDMs): LDM [79], SDXL [71], InstaFlow [56], SDXL-Turbo [86]
- Ablated diffusion: GLIDE [65]

Diffusion models with multi-stage pipelines include DALL-E2 [75], Wuerstchen [70], T-GATE [115], PIXART-α [8], PixArt-Σ [7], Ten [101], FouriScale [35], Make a Cheap Scaling [31], CogView3 [118], SDXL [71], and Frido [22]. The architectural improvements in diffusion models include:

- U-Net changes: FreeU [89], SCEdit [37]
- Mixture of Experts: ERNIE-ViLG 2.0 [24], eDiff-I [2], RAPHAEL [106]
- Hybrid models: YOSO [59], Diffusion2GAN [38], UFOGen [105]

Based on Prompt and Embedding, the diffusion models are:

- Vector quantization: VQ-Diffusion [30], Frido [22]
- Better text-image alignment: Imagen [82], Re-Imagen [10], Predicated Diffusion [94], CosmicMan [48], kNN-Diffusion [88]
- Prompt optimization: Prompt Expansion [14], DALL-E3 [4], CosmicMan [48], Parrot [47]

For training, the common strategies include:

- Distillation: SDXL-Turbo [86], SwiftBrush [64], InstaFlow [56], Diffusion2GAN [38]
- Neighbor-based methods: kNN-Diffusion [88], ConPreDiff [107]
- Shifted diffusion: Corgi [120]
- Disentanglement: PanGu-Draw [57]

Loss improvements appear in models like Self-perception [52], YOSO [59], Diffusion2GAN [38], and UFOGen [105]. Reinforcement learning is used in models like Parrot [47], RL-Diffusion [116], and RLCM [66]. LoRA is applied in Multi-LoRA Composition [119].

Their applications are as follows:

- Scene generation: SceneGraph2Image [61], Text2Street [93]
- Hand generation: HanDiffuser [62], Giving a Hand to Diffusion Model [69]
- Super-resolution: PIXART-α [8], PixArt-Σ [7], Ten [101], FouriScale [35], Make a Cheap Scaling [31], CogView3 [118]
- Other Models: The other models architecture used for text to image generation are: Energy-based models, PPGN [63] and Mamba-based models, ZigMa [34]
- Multimodal models are also used for image to text generation. These combine text, image, and other forms of data as inputs. Some of the models based on it are: Versatile Diffusion [104], GLIGEN [50], MiniGPT-5 [117], DiffusionGPT [72], RPG-DiffusionMaster [108], UNIMO-G [49], and CompAgent [102].

III.RESEARCH METHODOLOGY

In this section, a brief introduction of research methodology has been given. It explains how a machine learns to create images from text descriptions, like generating a flower image from a sentence, using a diffusion model and with attention mechanism and autoencoder model.

A. Diffusion Model

The entire process for methodology using diffusion model is performed using following steps:

1) The dataset comprised of flower images having a size of 64×64 each along with text descriptions. These texts were converted into embeddings, that were used to represent the meaning of the text. The images are were normalized into –1 to 1 for stability in the training. Data is fed to the model in small batches, such as 16 images at a time and shuffled for better learning.

2) The proposed model is a UNet based architecture with attention mechanism added so that it can better understand the text. It comprises of encoder and decoder components. The encoder, compress the image to perform feature extraction, while the decoder, decompress the latent features to recreate the image. Residual blocks were used in the architecture to help model train smoothly with a proper information flow and stabilize the training. The most important part in the architecture are Attention layers. They helped the image features “look at” the text embeddings to understand which visual details match the text.

3) The architecture follows diffusion technique to create image from text. Therefore, it takes a clean image as an input and slowly add noise to it until it becomes a pure noise. Later, in denoising process, the model learns, how to remove the noise added in steps. For that, it uses the text embeddings to get reference how a particular image looks according to the text.

4) The model is trained for 10 epochs with Adam optimizer and GPU for faster computation. Average losses were recorded for each training epochs.

The architecture for Diffusion Model is given as follows:

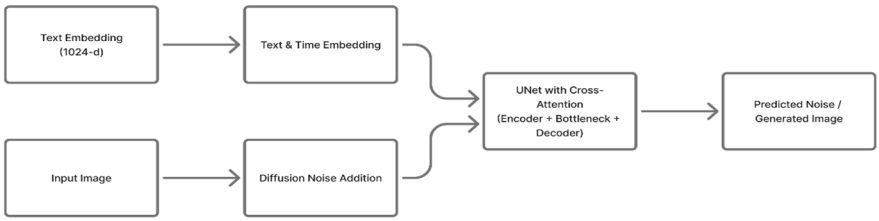


Fig. 1 Diffusion Model Architecture

B. Autoencoder

The entire process for methodology using autoencoder is performed using following steps:

1) Data Preparation The dataset is same as of used in diffusion model. For training stability, all images are normalized to the range [0, 1] by dividing pixel values by 255. A custom dataset class called Text2ImageDataset is implemented to load the dataset that includes the normalized RGB image, text embeddings, and the original text description. Only the training split of the dataset is utilized with a batch size of 32 and shuffling enabled for each epoch.

2) Model Architecture The proposed model is a conditional autoencoder with two primary components: an encoder and a decoder. The encoder progressively compresses the 64×64×3 input image using a sequence of convolutional layers with stride 2 for downsampling. It starts with 3 input channels and increases the feature depth to 64, then 128, and finally 256 channels, resulting in a compact latent representation of size 8×8×256. The text embedding (1024 dimensions) is first projected down to 256 dimensions using a fully connected (Linear) layer, then reshaped into a spatial feature map of size 8×8 and spatially expanded to match the latent dimensions. This text-conditioned feature map is concatenated channel-wise with the visual latent features, producing a combined latent representation of 512 channels. The decoder then upsamples this combined latent representation using a series of transposed convolutional layers (stride 2) to gradually increase the spatial size back to 64×64, reducing the channel count step-by-step from 512 to 256, then 128, then 64, and finally to 3 output channels. The final layer applies a 3×3 convolution followed by a Sigmoid activation to ensure the reconstructed image lies in the [0, 1] range. This simple encoder-decoder structure relies on direct channel concatenation for conditioning and does not include residual connections or attention mechanisms.

3) Training Procedure The model is trained for 10 epochs using the Adam optimizer with a learning rate of 1e-3. The average loss per epoch is accumulated and displayed, allowing observation of how reconstruction quality improves over time.

The architecture for Diffusion Model is given as follows:

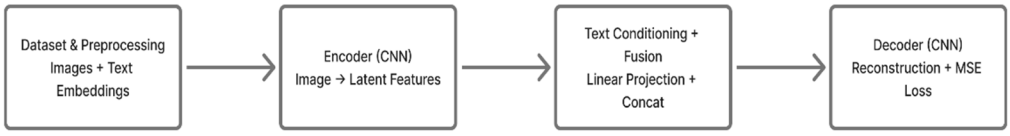


Fig. 2 Autoencoder Architecture

IV. COMPARATIVE RESULTS ANALYSIS: ATTENTION-BASED DIFFUSION VS CONDITIONAL AUTOENCODER

This section describes about the result analysis comparison for two models used for text to image generation tasks on the Flowers-HD5 dataset. Both models were trained for 10 epochs on GPU using MSE loss, but they solve different problems. The diffusion model focuses on image generation, while the autoencoder focuses on image reconstruction. The training performance and loss behavior comparison for both the models is given as follows:

A. Diffusion Model

1) Training starts with high loss because the model learns to remove strong noise.

2) During the first epoch, the average loss drops from about 0.95 to 0.35, showing fast learning.

3) Loss values fluctuate because random noise is added at every step.

4) Over 10 epochs, the loss is expected to stabilize around 0.1–0.2.

5) Training is slower because the model uses attention and multiple denoising steps.

B. Autoencoder Model

1) The autoencoder starts with much lower loss because it only reconstructs images.
2) Loss typically starts around 0.05 and reduces to about 0.01 by Epoch 10.
3) Training is stable and smooth, with very little fluctuation.
4) Training is faster due to a simpler architecture.

According to the training comparison, diffusion model achieved a higher initial loss but shows strong improvement. However, autoencoder converges faster, but learning may stop improving early. Both models reach low MSE values, but diffusion loss is more variable.

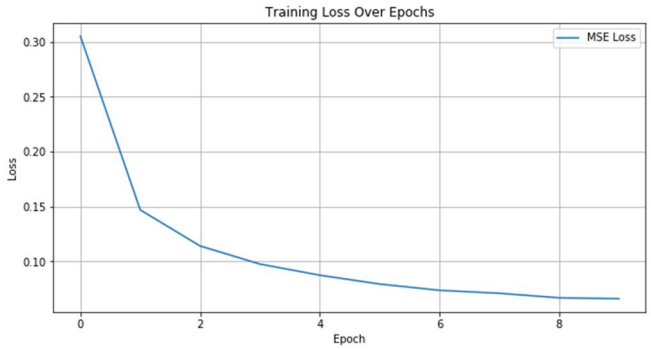


Fig 3. Training Loss over Epochs (Diffusion Model)

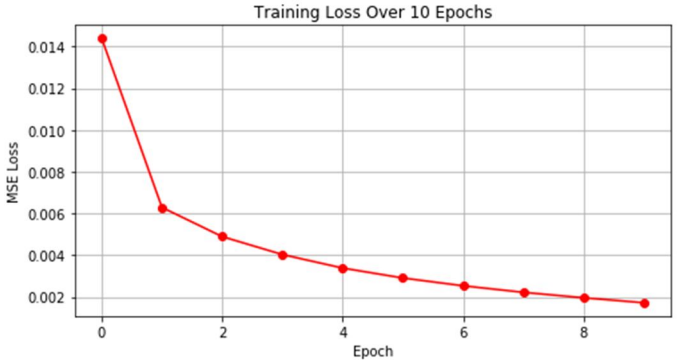


Fig 4. Training Loss over Epochs (Autoencoder Model)

The image quality and efficiency comparison for both the models is given as follows:

a) Diffusion Model Outputs

- Generates new and diverse flower images from text embeddings.
- Attention layers help match image details with text descriptions.
- Early outputs may show noise issues, but quality improves with training.
- Image generation is slow, as it requires many denoising steps.

b) Autoencoder Model Outputs

- Produces reconstructed versions of input images, not new images.
- Reconstructions are visually similar to the originals but slightly blurry.
- Text affects images only in a limited way.
- Very fast inference, since it is a single forward pass.

The diffusion model proved to be better for new image generation and semantic diversity. While, autoencoder is better for accurate reconstruction and speed. Diffusion benefits more from attention; autoencoder uses simple feature concatenation.

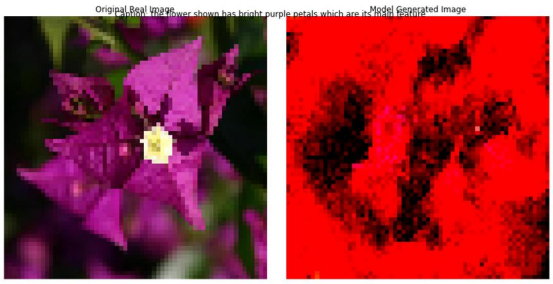


Fig 5. Diffusion Model Output

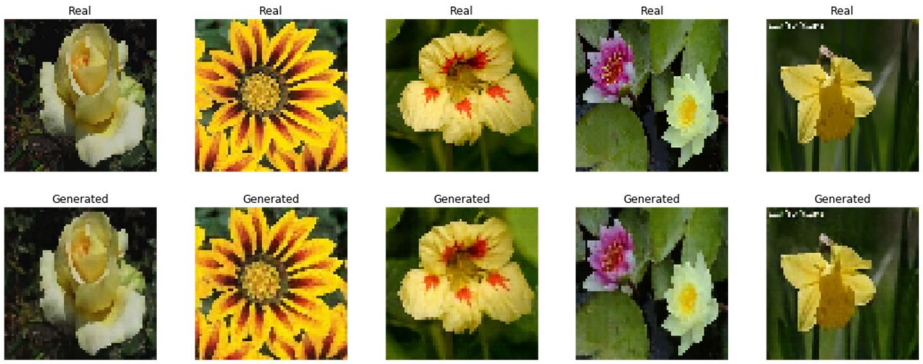


Fig 6. Autoencoder Model Output

V. CONCLUSION

This study compared a conditional autoencoder and an attention-based diffusion model for text-conditioned flower images generation. The autoencoder learned quickly and reconstructed images accurately with low error. But the kind of images generated were not new or diverse. The diffusion model trained more slowly and had higher loss, but it was able to generate completely new images that matched the text descriptions. Overall, the autoencoder is fast and efficient for reconstruction tasks, while the diffusion model is more powerful for text-to-image generation.

VI. FUTURE WORK

In future, we would like to use the autoencoder as a latent compressor for diffusion models. Also, we will add quantitative metrics such as CLIP score, PSNR for model evaluation. Training shall be performed on higher resolutions and more diverse datasets. Also, we will try to improve text conditioning in autoencoders using attention.

REFERENCES

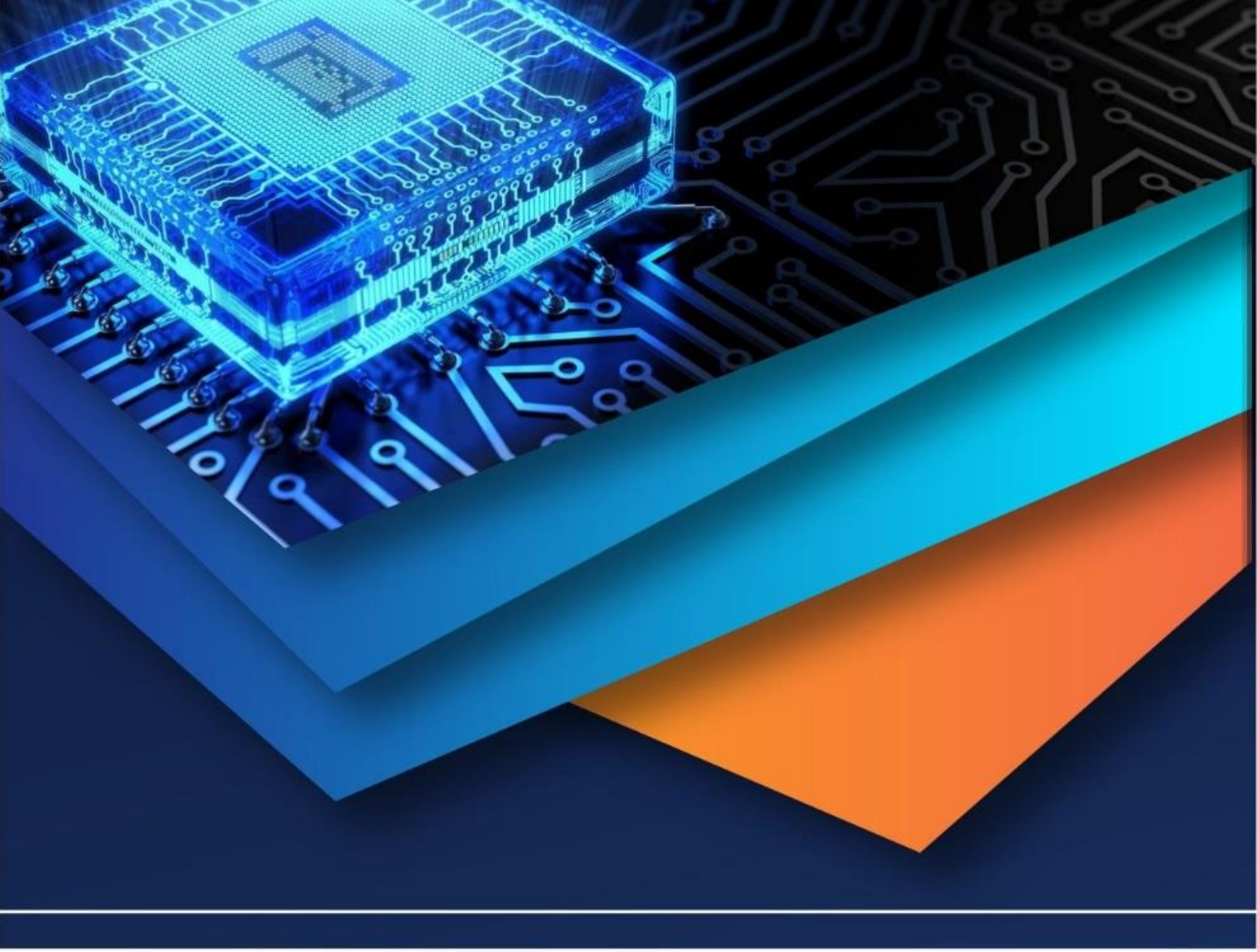
[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
[2] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, et al., "ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers," arXiv preprint arXiv:2211.01324, 2022.
[3] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," Advances in neural information processing systems, vol. 13, 2000.
[4] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al., "Improving image generation with better captions," Computer Science, 2023. https://cdn. openai. com/papers/dall-e-3. pdf 2, 3 (2023), 8.
[5] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M. Yang, K. Murphy, W. Freeman, M. Rubinstein, et al., "Muse: Text-to-image generation via masked generative transformers," arXiv preprint arXiv:2301.00704, 2023.
[6] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. Freeman, "Maskgit: Masked generative image transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11315–11325.
[7] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li, "Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation," arXiv preprint arXiv:2403.04692, 2024.
[8] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, et al., "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," arXiv preprint arXiv:2310.00426, 2023.
[9] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in International conference on machine learning, PMLR, 2020, pp. 1691–1703.
[10] W. Chen, H. Hu, C. Saharia, and W. Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," arXiv preprint arXiv:2209.14491, 2022.

[11] H. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., "Scaling instruction-finetuned language models," Journal of Machine Learning Research, vol. 25, no. 70, pp. 1–53, 2024.
[12] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. Bharath, "Generative adversarialnetworks: An overview," IEEE signal processing magazine, vol. 35, no. 1, pp. 53–65, 2018.
[13] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "Vqgan-clip: Open domain image generation and editing with natural language guidance," in European Conference on Computer Vision, Springer, 2022, pp. 88–105.
[14] S. Datta, A. Ku, D. Ramachandran, and P. Anderson, "Prompt expansion for adaptive text-to-image generation," arXiv preprint arXiv:2312.16720, 2023.
[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[16] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, et al., "Cogview: Mastering text-to-image generation via transformers," Advances in neural information processing systems, vol. 34, pp. 19822–19835, 2021.
[17] M. Ding, W. Zheng, W. Hong, and J. Tang, "Cogview2: Faster and better text-to-image generation via hierarchical transformers," Advances in Neural Information Processing Systems, vol. 35, pp. 16890–16902, 2022.
[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[19] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al., "Scaling rectified flow transformers for high-resolution image synthesis," in Forty-first International Conference on Machine Learning, 2024.
[20] P. Esser, R. Rombach, A. Blattmann, and B. Ommer, "Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis," Advances in neural information processing systems, vol. 34, pp. 3518–3532, 2021.
[21] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12873–12883.
[22] W. Fan, Y. Chen, D. Chen, Y. Cheng, L. Yuan, and Y. Wang, "Frido: Feature pyramid diffusion for complex scene image synthesis," in Proceedings of the AAAI conference on artificial intelligence, vol. 37, 2023, pp. 579–587.
[23] Z. Feng, R. Hu, L. Liu, F. Zhang, D. Tang, Y. Dai, X. Feng, J. Li, B. Qin, and S. Shi, "Emage: Non-Autoregressive Text-to-Image Generation," arXiv preprint arXiv:2312.14988, 2023.
[24] Z. Feng, Z. Zhang, X. Yu, Y. Fang, L. Li, X. Chen, Y. Lu, J. Liu, W. Yin, S. Feng, et al., "Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10135–10145.
[25] O. Gafni, A. Polyak, and Y. Taigman, "Scene-Based Text-to-Image Generation with Human Priors," US Patent App. 18/149,542, 2024.
[26] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer, "Mask-predict: Parallel decoding of conditional masked language models," arXiv preprint arXiv:1904.09324, 2019.
[27] R. Gray, "Vector quantization," IEEE Assp Magazine, vol. 1, no. 2, pp. 4–29, 1984.
[28] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in International conference on machine learning, PMLR, 2015, pp. 1462–1471.
[29] J. Gu, J. Bradbury, C. Xiong, V. Li, and R. Socher, "Non-autoregressiveneural machine translation," arXiv preprint arXiv:1711.02281, 2017.
[30] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10696–10706.
[31] L. Guo, Y. He, H. Chen, M. Xia, X. Cun, Y. Wang, S. Huang, Y. Zhang, X. Wang, Q. Chen, et al., "Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation," arXiv preprint arXiv:2402.10491, 2024.
[32] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.
[33] J. Ho and T. Salimans, "Classifier-free diffusion guidance," arXiv preprint arXiv:2207.12598, 2022.
[34] V. Hu, S. Baumann, M. Gui, O. Grebenkova, P. Ma, J. Fischer, and B. Ommer, "ZigMa: A DiT-style Zigzag Mamba Diffusion Model," arXiv preprint arXiv:2403.13802, 2024.
[35] L. Huang, R. Fang, A. Zhang, G. Song, S. Liu, Y. Liu, and H. Li, "FouriScale: A Frequency Perspective on Training-Free High-Resolution Image Synthesis," arXiv preprint arXiv:2403.12963, 2024.
[36] F. Jelinek, "Interpolated estimation of Markov source parameters from sparse data," in Proc. Workshop on Pattern Recognition in Practice, 1980.
[37] Z. Jiang, C. Mao, Y. Pan, Z. Han, and J. Zhang, "Scedit: Efficient and controllable image diffusion generation via skip connection editing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8995–9004.
[38] M. Kang, R. Zhang, C. Barnes, S. Paris, S. Kwak, J. Park, E. Shechtman, J. Zhu, and T. Park, "Distilling Diffusion Models into Conditional GANs," arXiv preprint arXiv:2405.05967, 2024.
[39] M. Kang, J. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up gans for text-to-image synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10124–10134.
[40] Y. Kim, C. Denton, L. Hoang, and A. Rush, "Structured attention networks," arXiv preprint arXiv:1702.00887, 2017.
[41] D. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.
[42] D. Kingma, M. Welling, et al., "An introduction to variational autoencoders," Foundations and Trends® in Machine Learning, vol. 12, no. 4, pp. 307–392, 2019.
[43] J. Koh, J. Baldridge, H. Lee, and Y. Yang, "Text-to-image generation grounded by fine-graineduser attention," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 237–246.
[44] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," arXiv preprint arXiv:1804.10959, 2018.
[45] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," arXiv preprint arXiv:1808.06226, 2018.

[46] D. Lee, C. Kim, S. Kim, M. Cho, and W. Han, "Autoregressive image generation using residual quantization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11523–11532.
[47] S. Lee, Y. Li, J. Ke, I. Yoo, H. Zhang, J. Yu, Q. Wang, F. Deng, G. Entis, J. He, et al., "Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation," arXiv preprint arXiv:2401.05675, 2024.
[48] S. Li, J. Fu, K. Liu, W. Wang, K. Lin, and W. Wu, "CosmicMan: A Text-to-Image Foundation Model for Humans," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6955–6965.
[49] W. Li, X. Xu, J. Liu, and X. Xiao, "UNIMO-G: Unified Image Generation through Multimodal Conditional Diffusion," arXiv preprint arXiv:2401.13388, 2024.
[50] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. Lee, "Gligen: Open-set grounded text-to-image generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22511–22521.
[51] J. Lin, R. Men, A. Yang, C. Zhou, M. Ding, Y. Zhang, P. Wang, A. Wang, L. Jiang, X. Jia, et al., "M6: A chinese multimodal pretrainer," arXiv preprint arXiv:2103.00823, 2021.
[52] S. Lin and X. Yang, "Diffusion Model with Perceptual Loss," arXiv preprint arXiv:2401.00110, 2023.
[53] L. Lingle, "Transformer-vq: Linear-time transformers via vector quantization," arXiv preprint arXiv:2309.16354, 2023.
[54] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," arXiv preprint arXiv:2209.03003, 2022.
[55] X. Liu, C. Gong, L. Wu, S. Zhang, H. Su, and Q. Liu, "Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization," arXiv preprint arXiv:2112.01573, 2021.
[56] X. Liu, X. Zhang, J. Ma, J. Peng, et al., "Instaflow: One step is enough for high-quality diffusion-based text-to-image generation," in The Twelfth International Conference on Learning Representations, 2023.
[57] G. Lu, Y. Guo, J. Han, M. Niu, Y. Zeng, S. Xu, Z. Huang, Z. Zhong, W. Zhang, and H. Xu, "PanGu-Draw: Advancing Resource-Efficient Text-to-Image Synthesis with Time-Decoupled Training and Reusable Coop-Diffusion," arXiv preprint arXiv:2312.16486, 2023.
[58] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," arXiv preprint arXiv:2310.04378, 2023.
[59] Y. Luo, X. Chen, and J. Tang, "You Only Sample Once: Taming One-Step Text-To-Image Synthesis by Self-Cooperative Diffusion GANs," arXiv preprint arXiv:2403.12931, 2024.
[60] E. Mansimov, E. Parisotto, J. Ba, and R. Salakhutdinov, "Generating images from captions with attention," arXiv preprint arXiv:1511.02793, 2015.
[61] R. Mishra and A. Subramanyam, "Scene Graph to Image Synthesis: Integrating CLIP Guidance with Graph Conditioning in Diffusion Models," arXiv preprint arXiv:2401.14111, 2024.
[62] S. Narasimhaswamy, U. Bhattacharya, X. Chen, I. Dasgupta, S. Mitra, and M. Hoai, "Handiffuser: Text-to-image generation with realistic hand appearances," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 2468–2479.
[63] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4467–4477.
[64] T. Nguyen and A. Tran, "Swiftbrush: One-step text-to-image diffusion model with variational score distillation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7807–7816.
[65] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," arXiv preprint arXiv:2112.10741, 2021.
[66] O. Oertell, J. Chang, Y. Zhang, K. Brantley, and W. Sun, "Rl for consistency models: Faster reward guided text-to-image generation," arXiv preprint arXiv:2404.03673, 2024.
[67] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in International conference on machine learning, PMLR, 2018, pp. 4055–4064.
[68] S. Patil, W. Berman, R. Rombach, and P. von Platen, "amused: An open muse reproduction," arXiv preprint arXiv:2401.01808, 2024.
[69] A. Pelykh, O. Sincan, and R. Bowden, "Giving a Hand to Diffusion Models: a Two-Stage Approach to Improving Conditional Human Image Generation," arXiv preprint arXiv:2403.10731, 2024.
[70] P. Pernias, D. Rampas, M. Richter, C. Pal, and M. Aubreville, "Würstchen: An efficient architecture for large-scale text-to-image diffusion models," arXiv preprint arXiv:2306.00637, 2023.
[71] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," arXiv preprint arXiv:2307.01952, 2023.
[72] J. Qin, J. Wu, W. Chen, Y. Ren, H. Li, H. Wu, X. Xiao, R. Wang, and S. Wen, "Diffusiongpt: LLM-driven text-to-image generation system," arXiv preprint arXiv:2401.10061, 2024.
[73] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning, PMLR, 2021, pp. 8748–8763.
[74] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of machine learning research, vol. 21, no. 140, pp. 1–67, 2020.
[75] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchicaltext-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, 2022.
[76] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in International conference on machine learning, PMLR, 2021, pp. 8821–8831.
[77] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," Advances in neural information processing systems, vol. 32, 2019.
[78] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in International conference on machine learning, PMLR, 2016, pp. 1060–1069.
[79] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[80] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer, 2015, pp. 234–241.
[81] S. Ruan, Y. Zhang, K. Zhang, Y. Fan, F. Tang, Q. Liu, and E. Chen, "Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 13960–13969.
[82] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, K. Ghasemipour, R. Lopes, B. Ayan, T. Salimans, et al., "Photorealistic text-to-image diffusion models with deep language understanding," Advances in neural information processing systems, vol. 35, pp. 36479–36494, 2022.
[83] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," Advances in neural information processing systems, vol. 29, 2016.
[84] S. Särkkä and A. Solin, "Applied stochastic differential equations," vol. 10, Cambridge University Press, 2019.
[85] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, "Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis," in International conference on machine learning, PMLR, 2023, pp. 30105–30118.
[86] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," arXiv preprint arXiv:2311.17042, 2023.
[87] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," arXiv preprint arXiv:1508.07909, 2015.
[88] S. Sheynin, O. Ashual, A. Polyak, U. Singer, O. Gafni, E. Nachmani, and Y. Taigman, "Knn-diffusion: Image generation via large-scale retrieval," arXiv preprint arXiv:2204.02849, 2022.
[89] C. Si, Z. Huang, Y. Jiang, and Z. Liu, "Freeu: Free lunch in diffusion u-net," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4733–4743.
[90] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," arXiv preprint arXiv:2010.02502, 2020.
[91] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," arXiv preprint arXiv:2303.01469, 2023.
[92] Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," arXiv preprint arXiv:2011.13456, 2020.
[93] J. Su, S. Gu, Y. Duan, X. Chen, and J. Luo, "Text2Street: Controllable Text-to-image Generation for Street Views," arXiv preprint arXiv:2402.04504, 2024.
[94] K. Sueyoshi and T. Matsubara, "Predicated Diffusion: Predicate Logic-Based Attention Guidance for Text-to-Image Diffusion Models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8651–8660.
[95] M. Tao, H. Tang, F. Wu, X. Jing, B. Bao, and C. Xu, "Df-gan: A simple and effective baseline for text-to-image synthesis," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16515–16525.
[96] A. Vahdat, E. Andriyash, and W. Macready, "Dvae#: Discrete variational autoencoders with relaxed boltzmann priors," Advances in Neural Information Processing Systems, vol. 31, 2018.
[97] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al., "Conditional image generation with pixelcnn decoders," Advances in neural information processing systems, vol. 29, 2016.
[98] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in International conference on machine learning, PMLR, 2016, pp. 1747–1756.
[99] A. Van Den Oord, O. Vinyals, et al., "Neural discrete representation learning," Advances in neural information processing systems, vol. 30, 2017.
[100] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
[101] X. Wang, J. Kontkanen, B. Curless, S. Seitz, I. Kemelmacher-Shlizerman, B. Mildenhall, P. Srinivasan, D. Verbin, and A. Holynski, "Generative powers of ten," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7173–7182.
[102] Z. Wang, E. Xie, A. Li, Z. Wang, X. Liu, and Z. Li, "Divide and conquer: Language models can plan and self-correct for compositional text-to-image generation," arXiv preprint arXiv:2401.15688, 2024.
[103] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, and N. Duan, "Nüwa: Visual synthesis pre-training for neural visual world creation," in European conference on computer vision, Springer, 2022, pp. 720–736.
[104] X. Xu, Z. Wang, G. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7754–7765.
[105] Y. Xu, Y. Zhao, Z. Xiao, and T. Hou, "Ufogen: You forward once large scale text-to-image generation via diffusion gans," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8196–8206.
[106] Z. Xue, G. Song, Q. Guo, B. Liu, Z. Zong, Y. Liu, and P. Luo, "Raphael: Text-to-image generation via large mixture of diffusion paths," Advances in Neural Information Processing Systems, vol. 36, 2024.
[107] L. Yang, J. Liu, S. Hong, Z. Zhang, Z. Huang, Z. Cai, W. Zhang, and B. Cui, "Improving diffusion-based image synthesis with context prediction," Advances in Neural Information Processing Systems, vol. 36, 2024.
[108] L. Yang, Z. Yu, C. Meng, M. Xu, S. Ermon, and C. Bin, "Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms," in Forty-first International Conference on Machine Learning, 2024.
[109] Y. Yang, L. Wang, D. Xie, C. Deng, and D. Tao, "Multi-sentence auxiliary adversarial networks for fine-grained text-to-image synthesis," IEEE Transactions on Image Processing, vol. 30, pp. 2798–2809, 2021.
[110] J. Yu, X. Li, J. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized image modeling with improved vqgan," arXiv preprint arXiv:2110.04627, 2021.
[111] J. Yu, Y. Xu, J. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, et al., "Scaling autoregressive models for content-rich text-to-image generation," arXiv preprint arXiv:2206.10789, 2022.
[112] H. Zhang, J. Koh, J. Baldridge, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 833–842.
[113] H. Zhang, W. Yin, Y. Fang, L. Li, B. Duan, Z. Wu, Y. Sun, H. Tian, H. Wu, and H. Wang, "Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation," arXiv preprint arXiv:2112.15283, 2021.

[114] J. Zhang, Y. Han, P. Zhang, J. Yang, L. Zhang, J. Gao, P. Wang, and L. Yuan, "LAFITE: Towards Language-Free Training for Text-to-Image Generation," arXiv preprint arXiv:2111.13792, 2021.
[115] W. Zhang, H. Liu, J. Xie, F. Faccio, M. Shou, and J. Schmidhuber, "Cross-attention makes inference cumbersome in text-to-image diffusion models," arXiv preprint arXiv:2404.02747, 2024.
[116] Y. Zhang, E. Tzeng, Y. Du, and D. Kislyuk, "Large-scale Reinforcement Learning for Diffusion Models," arXiv preprint arXiv:2401.12244, 2024.
[117] K. Zheng, X. He, and X. Wang, "Minigpt-5: Interleaved vision-and-language generation via generative vokens," arXiv preprint arXiv:2310.02239, 2023.
[118] W. Zheng, J. Teng, Z. Yang, W. Wang, J. Chen, X. Gu, Y. Dong, M. Ding, and J. Tang, "Cogview3: Finer and faster text-to-image generation via relay diffusion," arXiv preprint arXiv:2403.05121, 2024.
[119] M. Zhong, Y. Shen, S. Wang, Y. Lu, Y. Jiao, S. Ouyang, D. Yu, J. Han, and W. Chen, "Multi-lora composition for image generation," arXiv preprint arXiv:2402.16843, 2024.
[120] Y. Zhou, B. Liu, Y. Zhu, X. Yang, C. Chen, and J. Xu, "Shifted diffusion for text-to-image generation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 10157–10166.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY