Research On Text Generated Images Based on GAN And Diffusion

Junduo Zheng

School of mathematics and statistics, Northeastern University, Qinhuangdao, China

202215079@stu.neu.edu.cn

Abstract. With the continuous development of deep learning, the generation of content by artificial intelligence has become a hot topic. Especially in the field of text to image generation, significant progress has been made. This article comprehensively compares text image generation methods based on Generative Adversarial Network and Diffsuion and their applications in text image generation tasks, demonstrating their respective advantages and limitations as well as possible solutions. Meanwhile, this article delves into the specific methods of diffusion models in improving image quality, optimizing model efficiency, and generating images based on multilingual text prompts. Through experimental analysis on the Microsoft Common Object Context Dataset (COCO), the zero sample generation ability of the diffusion model and the performance improvement of the GAN model were verified, highlighting the advantages of the GAN model in terms of lightweight and small parameter count. In addition, this article introduces the broad application prospects of diffusion models in fields such as text 3D generation and video generation. Finally, the challenges and future development trends faced by diffusion models in text to image generation tasks were summarized, with the hope of providing convenience for further research in this field.

Keywords: GAN, Diffusion, Text to Image, Image Multimodal.

1. Introduction

Text generated image technology refers to the technique of converting text descriptions into images through computer programs. Users input a text description, and the system generates corresponding images based on this description to achieve the mapping from text to image.

With the continuous development of AIGC, image generation has been applied in many different fields, from artistic creation to virtual anchors, and text generated image applications are everywhere in life.

Text image generation technology has shown broad application prospects in practical industries and has enormous development potential.

Firstly, from the perspective of cost and efficiency, this technology can greatly shorten the production cycle. Taking the advertising design industry as an example, in the past, it may take several days for designers to manually draw advertising posters. With the help of text image generation technology, corresponding text descriptions can be input according to requirements, and preliminary design schemes can be generated in just a few minutes, greatly reducing the production cycle. At the same time, it reduces labor costs, and enterprises do not need to hire a large number of professional graphic designers. Only a few personnel familiar with technical operations can complete most of the image generation work. The work efficiency has also been significantly improved, able to quickly respond to market demand and timely launch image products that meet requirements.

Secondly, in terms of innovation and flexibility, text image generation technology has brought more flexible, innovative, and efficient solutions to various industries. In game development, developers can quickly generate different styles of game scenes and character images according to the plot requirements, making it easy to implement creative ideas. In the field of architectural design, designers can generate images of various building appearances and interior layouts by inputting text descriptions, providing more possibilities for design.

Among them, the development of text generated image technology has gone through a process of refinement from traditional methods to deep learning, and then to multimodal large models.

According to different technical frameworks, the development of text generation image modeling technology has gone through three main stages: GAN, AR, and Diffusion. In the in-depth exploration of the field of text image generation (Raffaella et al, 2016), GAN based methods and Diffusion based methods have been separately and comprehensively sorted and classified (Li et al, 2025), laying a solid foundation for research and application in this field. However, the review, comparison, and analysis of these two methods in text image generation tasks are still an urgent research area that needs to be improved. This review will focus on introducing the basic methods and improvement methods of text generated images based on GAN and fusion, analyzing the advantages and disadvantages of each method, the problems currently encountered, and possible solutions. (Mohamed, Omar, Somaya et al，2022)

2. Text generated images based on GAN

2.1. Basic Principles of Gan Framework

The task of image generation is mainly based on keyword or sentence retrieval to register visual content that matches text. With the emergence of generative adversarial networks, text to image synthesis methods have made significant progress in visual realism.

GAN is divided into two parts: generator and discriminator. The generator generates an image by accepting a random noise, with the aim of deceiving the discriminator as much as possible. The discriminator receives an image and outputs a probability of 0-1 to indicate the probability that the image is real. Figure 1 shows an overview of text generated images based on GAN. During the training process, the generator continuously improves the authenticity of the generated images, and the discriminator continuously improves its accuracy in distinguishing authenticity, ultimately training a generator that can generate high-quality data and a discriminator that can accurately distinguish the authenticity of data. When training the generator, the discriminator performance needs to be fixed;When training the discriminator, the performance of the generator needs to be fixed. By alternating between training the generator and discriminator, the performance of both can be improved. The generator plays a significant role in this context. The recognizer undergoes training to differentiate between genuine and fabricated data. As training time elapses, the accuracy of the recognizer steadily rises. As the training process advances, both the generator network and the recognizer network engage in an adversarial process. The main objective of the generator is to create samples that are indistinguishable from real world data by the recognizer. On the other hand, the recognizer makes every effort to precisely tell apart real data from the fake samples produced by the generator. During this adversarial interaction, the performance of both the generator and the discriminator gets enhanced. Eventually, the generator reaches a stage where it can generate images whose distribution closely resembles that of real world data. (Li, Tong, Zhao et al, 2023).

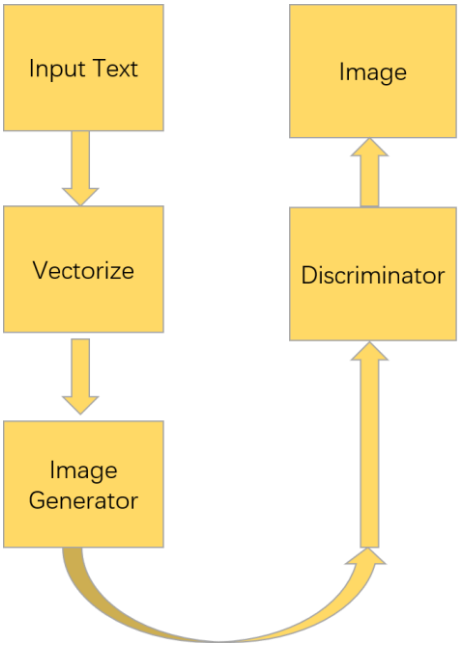

Figure 1: GAN Based Image Generation General (Picture credit: Original)

The generator loss represents the extent to which the generator can deceive pseudo-random numbers. When the samples generated by the generator are used as input, the calculation is carried out based on the output of pseudo random numbers. The generator aims to achieve the minimization of this loss. By doing so, it is impelled to generate samples that are more in line with the characteristics of real world data. Therefore, the optimization process of GAN can be summarized as a binary minimax game problem, and the network loss function is:

$$\min_{G}\max_{D} V(D,G) = \min_{G}\max_{D} \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

In this formula, $p_{data}(x)$ is the true data distribution; z is random noise; p_z is the noise distribution; G is the generation mapping function; D is the identification mapping function.

2.2. GANS Variants for Text Generated Images: AttnGAN

A prevalent approach involves encoding the whole textual description into a global sentence level vector. This vector then serves as a conditional factor for the image generation process based on neural networks. Specifically, it is employed to fine tune the Generative Adversarial Network (GAN) in the context of the global sentence vector. But at the word level, there is often a lack of important fine-grained information. To solve this problem, AttnGAN emerged. AttnGAN can synthesize fine-grained details in different subregions of an image by focusing on relevant words in natural language descriptions.

This model has added two important components: AttnGAN and DAMSM

Among the techniques, AttnGAN transforms natural language descriptions to form a global sentence vector. Additionally, each individual word in the sentence gets encoded into a word specific vector. Generate low resolution images using global sentence vectors in the first stage of the generative network. In the subsequent steps, it takes the image vectors from every sub region and employs an attention layer to construct word context vectors and query word vectors. After that, the model fuses the region image vector with the corresponding word context vector to create a multimodal context vector. Based on this multimodal context vector, it generates new image features in the neighboring sub regions. This process effectively generates higher resolution images with more details at each stage.

By leveraging the attention mechanism, DAMSM is capable of utilizing both global sentence level information and fine grained word level information to compute the similarity between the generated

images and sentences. As a result, DAMSM offers an additional fine grained image text matching loss for training the generators.

In scenarios that require fine-grained operations, such as medical imaging lesion analysis, this method significantly improves detail restoration through subunit level feature extraction; However, tasks involving large-scale spatial layout or complex logical associations, such as panoramic image generation, still require strengthened global consistency constraint mechanisms.

2.3. The GANS Variant of This Generated Image: MirrorGAN

MirrorGAN adopts the concept of learning text to image generation via re description and is composed of three main modules. Semantic Text Embedding Module (STEM) is designed to generate both word level and sentence level embeddings. Global Local Collaborative Attention Module (GLAM) features a cascaded architecture for image generation. It starts from generating a coarse scale version of the target image and then refines it step by step to a fine scale image. Semantic Text Regeneration and Alignment Module (TREAM) plays a role in regenerating the semantic text and aligning it with the generated images.

The mirror mechanism is the core innovation of MirrorGAN. Its primary function is to capture both global and local details in the generated images.

Taking a city landscape map as an example, global details may include the overall layout of the city, the outline of the skyline, etc;Local details include shop signs on the street and the clothing of pedestrians. The mirror mechanism can be like a meticulous photographer, able to grasp the whole picture of the scene and focus on every tiny element (Qiao, Zhang, Xu et al, 2019).

The multi - level mirroring mechanism empowers the generator to fine tune and synchronize the generated images across various scales. It does so by taking the images produced at higher levels as benchmarks for the generation process at lower levels. This is like building a skyscraper, first building the overall framework (high-level generated images), and then gradually decorating and arranging each floor and room based on this framework (low-level generation process). In the process of image generation, through this approach, the generator can make the image more coordinated at various scales, resulting in visually more coherent, high-quality, detail rich, and context relevant images (Sibi, 2024).

3. Wensheng Graph Based on Stable Diffusion

3.1. Basic Principles of Diffusion Framework

Diffusion is a generative model based on probability diffusion process. The uniqueness of this method lies in its adoption of a unique bottom-up modeling approach. It does not directly create images from textual descriptions, but rather utilizes a gradually unfolding program that starts from completely random noise, gradually adds structure and details, and ultimately forms a meaningful image. The Diffusion model is mainly divided into two basic steps. Among them, the forward diffusion process iteratively applies noise to the image until a completely noisy image is generated; Backward denoising process: Input random noise (image size), use Unet network to predict the noise added in the previous step, output the denoising result of the previous step, and finally output a generated image that conforms to the probability distribution (Jonathan, Ajay, Piete, 2020). The mainstream diffusion methods include Latent Diffusion model and Stable Diffusion model.

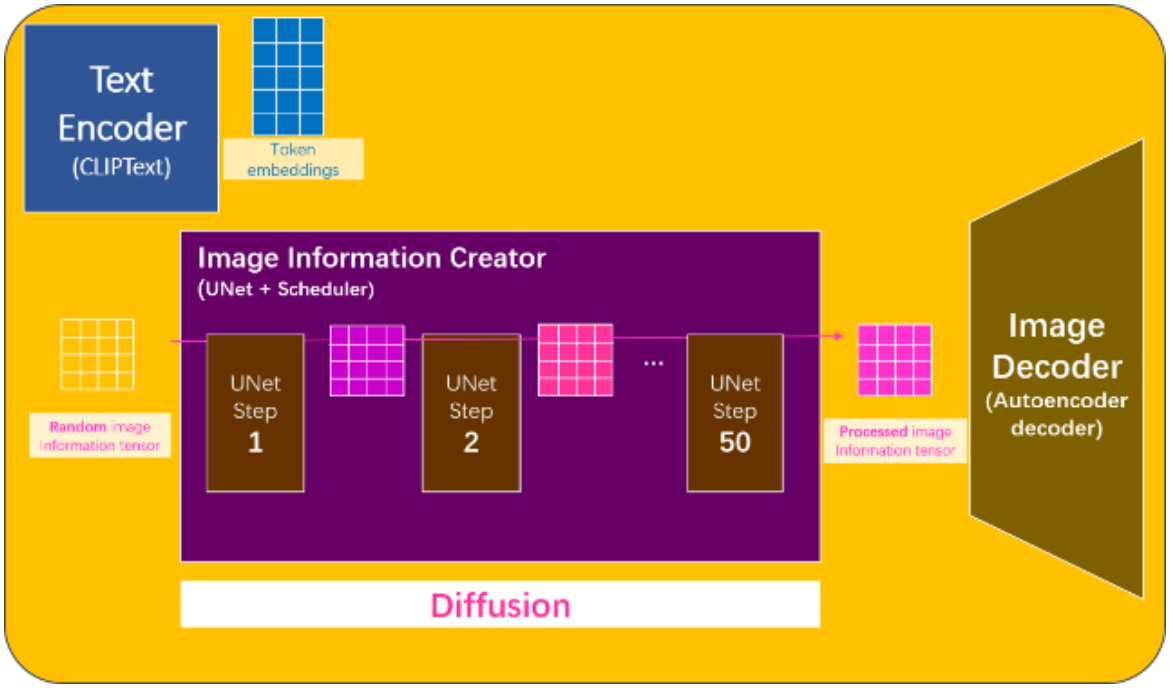

Figure 2: Stable Diffusion (Picture credit: Original)

Figure 2 shows the structure of a stable diffusion model. Diffusion models (DMs) work directly in the pixel domain, and optimization and inference are time-consuming. In order to train them on limited computing resources, LDM first uses a pre trained AutoEncoder to convert image pixels to smaller dimensional latent spaces, significantly reducing the computational complexity of the diffusion process. Then proceed with traditional diffusion model inference and optimization. This training method strikes a balance between computing power and performance for LDM. In addition, by introducing cross attention, DMs can achieve good results in conditional generation, including text generated images (Liu, Yin, Xue et al，2023).

3.2. Improvement of Text Generated Images Based on Diffusion Framework: Stable Diffusion

Stable Diffusion is an open-source text to image generation model built on the LDM framework. It optimizes text conditional generation, integrates pre trained models such as CLIP, and supports multimodal input (such as text+sketch generated images).

Stable Diffusion consists of three key components: clip, VAE, and U-Net network.

Clip serves as a "bridge" for the conversion between text information and machine data information. As a front-end module, it encodes the input text information to generate a Text Embeddings feature matrix corresponding to the text information, and then controls the generation of the image using the input of t. The CLIP model is a multimodal model based on contrastive learning, mainly consisting of two models: Text Encoder and Image Encoder. The Text Encoder is used to extract text features, such as the Text Transformer model; Image Encoder is mainly used to extract features of images, such as ResNet and ViT. At the same time, clip directly trains the dataset with 400 million images and labeled text to learn the correspondence between images and the content of this article. The Text Encoder section is mainly used in Stable Diffusion. The CLIP Text Encoder model encodes the input text Prompt into Text Embeddings (semantic information of the text), which are embedded into Stable Diffusion through the CrossAttention module of the U-Net network as Conditions to control and guide the content of the generated image to a certain extent. Text Embeddings are used as U-Net. Currently, the Text Encoder model used in Stable Diffusion is CLIP ViT-L/14CLIP ViT-L/14.

Variational Auto Encoder (VAE) is a generative model based on the Encoder Decoder architecture. The Encoder structure of VAE can convert input images into low dimensional Latent features and use them as inputs for U-Net. The decoder structure of VAE can reconstruct low dimensional Latent

features to restore pixel level images. Overall, in Stable Diffusion, the VAE model mainly plays a role in image compression and image reconstruction. There are three basic components in the VAE model: GSC component, Downsample component, and Upsample component. And two core components: ResNetBlock module and SelfAttention model.

Net is the core module in Stable Diffusion. U-Net mainly iteratively denoises Gaussian noise matrices in the "diffusion" loop, and each predicted noise is guided by text and timesteps to remove the predicted noise on a random Gaussian noise matrix, ultimately transforming the random Gaussian noise matrix into hidden features of the image. U-Net in Stable Diffusion adds Time Embedding module, Cross Attention module, and Self Attention module on the basis of Encoder Decoder structure. During the diffusion loop of U-Net, the Content Embedding module remains unchanged while the Time Embedding module changes every time. Each time the noise predicted by U-Net is subtracted from the Latent feature, and the iterated Latent is used as the new input for U-Net. In Stable Diffusion, the Cross Attention module is used to control the fusion and interaction of text and image information, and U-Net is controlled to associate a certain block of the noise matrix with specific information in the text.

3.3. Latest Improvement in Text Generated Images Based on the Diffusion Framework: Stable Diffusion 3

Stable Diffusion 3 is the latest and most powerful text to image model from Stability AI. It has significant improvements in handling multi topic prompts, image quality, and even text rendering capabilities. The most significant difference is that it combines DiT architecture and FM architecture. DiT is particularly effective in class conditional image generation tasks on large datasets such as ImageNet, setting new benchmarks in image quality and generation model performance;FM not only makes the training of diffusion models more robust, but also provides faster training, sampling, and better generalization ability for CNFs that use non diffusion probability paths (such as optimal transmission paths).

Common datasets for text generated images

CUB-200-2011, Oxford-102 Flower, MSCOCO, COCO

The CUB-200-2011 dataset is an extended version of CUB-200, which is a dataset of 200 bird species. The extended version roughly doubled the number of images for each category and added new part positioning annotations. All images are annotated with bounding boxes, part positions, and attribute labels.

Oxford-102 Flower is a flower dataset for image classification released by the Oxford University of Engineering in 2008, with selected flowers typically located in the UK.

The full name of the MSCOCO dataset is Microsoft Common Objects in Context. It is a large image dataset developed and maintained by Microsoft, with tasks including recognition, segmentation, and detection.

COCO is a highly regarded and large-scale dataset used in scenarios such as object detection, segmentation, image description, and more.

To evaluate the performance of GAN and diffusion models in more challenging scenarios, table 1 presents the Supervised FID and Zero shot FID values of some algorithms on the MSCOCO dataset.

The Supervised Frechet Inception Distance (FID) metric serves to assess the resemblance between the distribution of images generated under supervision and that of real world image data. In contrast, the Zero shot FID metric is employed to evaluate the similarity between the distribution of unsupervised generated images and actual real image data (Xu, Chen, 2025).

Table 1: Comparison of the effectiveness of different text generated image methods on the COCO dataset

year	model	type	Supervised FID	Zero - shot FID	Parameter quantity / 10^8
2018	AttnGAN	GAN	35.49	—	—
2019	DM - GAN	GAN	32.64	—	—
2019	DF - GAN	GAN	21.42	—	0.019
2019	SDGAN	GAN	29.35	—	—
2021	DAE - GAN	GAN	28.12	—	—
2021	XMC - GAN	GAN	9.33	—	—
2022	LAFITE	GAN	8.12	26.94	0.15
2023	GALIP	GAN	12.54	—	0.24
2021	GLIDE	DM	—	12.24	5
2022	DALL·E 2	DM	—	10.39	6.5
2022	VQ - Diffusion	DM	—	13.86	0.37
2022	Improved VQ - Diffusion	DM	—	8.44	1.27
2022	LDM	DM	—	12.63	1.45
2022	Stable v1.4	DM	—	16.31	0.9
2022	Imagen	DM	—	7.27	7.9
2023	Swinv2 - Imagen	DM	—	7.21	—
2022	Upainting	DM	—	8.34	—
2022	Corgi	DM	—	10.6	—
2022	eDiff - I^2	DM	—	6.95	9.1
2023	ERNIE - ViLG 2.0	DM	—	6.75	24
2023	RAPHAEL	DM	—	6.61	3
2023	PixArt - α	DM	—	7.32	0.6
2023	PanGu - Draw	DM	—	7.99	5
2022	Re - Imagen	DM	5.25	6.88	3.6
2023	Kandinsky	DM	—	8.03	3.3
2023	SDXL	DM	—	11.93	2.6
2023	UFOGen	DM	—	12.78	0.9
2022	Knn - Diffusion	DM	—	12.5	0.4
2023	BK - SDM - Small	DM	—	13.57	0.76

According to the experimental data in the table 1, with the gradual improvement of the GAN (GAN) research system, the model shows a gradually improving performance in the Supervised FID index. This trend indicates that as the model optimization strategy is iteratively updated, the generator can more accurately capture and learn the potential distribution features of the target dataset, while generating image samples that are highly consistent with the original data distribution, resulting in a significant improvement in generation quality. Among them, LAFITE even reached 8.12. However, the GAN model has limitations in overly relying on the training distribution, and its generation ability is limited when facing zero sample scenarios. The LAFITE value on Zero shot FID is only 26.94 (Gao, Du, Song, 2024).

The Diffusion model gradually restores the image through denoising steps, which allows the model to creatively generate images during the generation process. Therefore, it can also generate high-quality images on unseen data. The quantitative analysis based on the table 1 shows that the diffusion model exhibits breakthrough zero sample generation performance in the COCO-30K benchmark test, mainly due to its synergistic effects in three dimensions: cross modal semantic parsing ability, network topology innovation, and joint optimization of training strategies. This type of model not only achieves high-quality visual output in the FID metric dimension (such as Imagen reaching 7.27), but more importantly, effectively maintains the generalization characteristics of open domain concepts through probability density matching mechanism. Subsequent research should focus on enhancing the generalization robustness of models in unknown concepts and open scenarios,

especially exploring the technical path of combining multimodal alignment mechanisms with dynamic adaptive architectures.

The parameter count, as a key indicator of model complexity, can reflect the learning and expression abilities of the model, and is closely related to the demand for computing resources and training efficiency. GAN (GAN) constructs a dynamic game optimization framework through adversarial training mechanism, which can achieve implicit modeling of data distribution without relying on large-scale parameter quantities. This architectural innovation has been validated in models such as DF-GAN (1.9×10^{7}), GALIP (2.4×10^{8}), and LAFITE (1.5×10^{8}):Thanks to its unique parameter efficiency advantage, by implicitly modeling the core feature extraction ability of the data manifold structure, compared to the parameter size of traditional generation architectures (usually exceeding 109), this type of model significantly optimizes the memory occupancy during gradient backpropagation while ensuring generation quality. This lightweight architecture design effectively controls computational complexity, making it particularly suitable for image generation tasks in resource constrained scenarios.

The diffusion model adopts a method of gradually eliminating noise to restore the image, focusing on the iterative process of noise prediction and elimination. The parameter size of this model ranges from 3.7×10^{8} to 2.4×10^{10}, which reflects the variability of the diffusion model in terms of parameter size and its good adaptability to various application scenarios.

4. Discussion

In the GAN text generation graph model, the main challenges are as follows. (1) Lightweight models face the technical bottleneck of insufficient cross domain generalization ability in complex scene image synthesis tasks, and their generated results have significant gaps with human cognitive standards in dimensions such as semantic coherence (such as multi object spatial relationship accuracy below 60%) and visual rationality (FID values generally above 40).

The current implementation of human level image generation quality (such as FID<5) relies on giant generation architectures with parameter scales exceeding 1E+11 (such as DALL · E3, Stable Diffusion XL, etc.). The training of such models requires tens of thousands of GPU hours of computing resources (with a single training cost exceeding $3 million), and is limited by the difficulty of building and maintaining a distributed training cluster at the kilocard level. Only top companies such as Meta and OpenAI have a complete technology chain globally.

The very large-scale generation model has the core defect that the single reasoning delay is too high (for example, the reasoning time of Stable Diffusion on the A100 graphics card is>15 seconds). Even through model quantification (the memory occupation is still more than 10GB under the 8bit precision) and knowledge distillation and other technical optimization, it is still difficult to meet the deployment requirements of mobile terminals (for example, the end-to-end reasoning delay of smart phones is less than 1 second) and edge computing devices (the memory capacity is usually less than 8GB), which seriously restricts the landing of industrial applications.

To solve problem one, the cross domain knowledge of the multi billion parameter model can be transferred in stages to a lightweight model, and a multi-layer semantic distillation mechanism can be designed. Build a knowledge transfer tree consisting of global semantics (scene topology) - mid-level features (object associations) - low-level textures (material details), and dynamically select valuable feature layers (such as object interaction features) through attention gating algorithms.

To solve problem two, a distributed training strategy combining model parallelism and data parallelism can be adopted, and mixed precision training techniques can be used. Model parallelism allocates different parts of the model to different computing devices for computation, while data parallelism allocates different data samples to different devices for training. The combination of the two can fully utilize the computing resources of multiple computing devices and improve training efficiency. Mixed precision training uses semi precision floating-point numbers (such as FP16) for computation, which can reduce memory usage and computational complexity, thereby lowering

training costs. For example, when training a giant generative architecture, assigning different layers of the model to different GPUs, while assigning different image data samples to each GPU for training, and using FP16 for computation, can significantly improve training speed without sacrificing too much accuracy.

To solve the third problem, we can develop model compression and acceleration technologies specifically for mobile terminals and edge computing devices, such as combination optimization of pruning, low rank decomposition and model quantification. Pruning can remove neurons and connections in the model that have little impact on inference results, reducing the number of model parameters; Low rank decomposition can decompose large matrices into the product of multiple small matrices, reducing computational complexity; Model quantization can convert the parameters and calculations of the model from high-precision floating-point numbers to low precision integers, reducing memory usage and computational complexity. Through the combination and optimization of these three technologies, the reasoning delay of the model can be significantly reduced, so that it can run quickly on mobile terminals and edge computing devices. For example, pruning a large-scale generative model to remove some redundant neurons, optimizing the convolutional layers using low rank decomposition, and finally quantifying the model into 8-bit or 4-bit integers can greatly reduce the computational complexity and memory usage of the model.

Currently, there are also some issues with the diffusion model.

The quality issues, lack of diversity, and imbalanced multilingual data in datasets are becoming key factors that constrain the further improvement and widespread application of model performance. In many datasets, there is a deviation between annotated information and actual image content. The dataset may contain a large amount of noise data that is unrelated to the target task. There is a lack of consistency in the presentation of data within the same category. The dataset covers a limited range of scenes, with a single scene and a single image style. And datasets often lack samples with special circumstances.

The existing diffusion models suffer from problems such as a large number of model parameters, high consumption of computational resources, and low sampling efficiency, which seriously affect the practical application results.

The evaluation method is too one-sided and subjective. The evaluation of current text to image generation methods mainly relies on specific metrics or manual evaluation, but these methods have certain limitations. If FID value is used as an evaluation indicator, it is not always consistent with the perceived quality of the image, and manual evaluation is limited by aesthetic differences between evaluators and inefficient.

Regarding problem one. Efficiently identify and eliminate bad data by combining machine learning algorithms with manual review. At the same time, it is possible to explore the approach of text association learning without language environment. Traditional text image association learning usually relies on language annotation, but can discover potential associations between text and images by mining visual features and patterns in images and using unsupervised or self supervised learning algorithms. Enhance data using external knowledge graphs. The external knowledge graph contains rich knowledge and information, which can provide a broader background and context for text image datasets. By combining knowledge graphs with datasets, the depth and breadth of data can be enhanced.

Regarding problem two, the following methods can be used to solve it. Shorten the length of the diffusion chain. By using mathematical deduction methods, we delve into the intrinsic mechanisms of diffusion processes and find effective ways to shorten the length of diffusion chains. For example, by establishing mathematical models, analyzing the relationships between various steps in the diffusion process, and identifying steps that can be simplified or merged, unnecessary calculations can be reduced and training efficiency can be improved. Reduce model parameters: Distillation learning is employed as a mechanism to transfer the knowledge embedded within large - scale teacher models to smaller student models. This can significantly reduce the number of model parameters while maintaining model performance. To accelerate the sampling speed and reduce the complexity

of model training, some new high-quality sampling strategies can be proposed. For example, differential equation solving sampler, dynamic programming sampling, implicit sampling DDIM.

To address problem three, a unified evaluation framework can be constructed to enhance the universality and generalization ability of text to image generation models. This framework has clear and diverse evaluation indicators, which can comprehensively and objectively measure the performance of the model, making it easy for researchers to make fair comparisons.

5. Conclusion

The text generation image task based on GAN involves many classic GAN models. With the gradual improvement of the research system of GAN, the quality of image generation in text generation image tasks has significantly improved with the improvement of GAN based models. Models can more accurately understand and learn the distribution of real images, generate images that are highly consistent with real images, and have generally small parameter quantities, low learning and training costs. They are particularly suitable for text generation image tasks in resource limited scenarios.

The text generation image model based on Difusion allows the model to creatively generate images during the generation process. It can also generate high-quality and semantically related images in zero sample generation tasks. It has a variety of improved models, including low-cost models with small parameter quantities and high-precision models with high parameter quantities. In practical applications, model selection can be made based on specific task requirements and computational resource limitations.

The text generation image system, with the help of deep learning technology and trained on massive amounts of data, has been able to parse text descriptions to a certain extent and generate corresponding images based on them. The advancement of this technology has opened up new development space for the field of artificial intelligence, especially in the creative industry, showing extremely promising application prospects Overall, this technology is still in the process of growth, but its potential application scope is extremely broad. The method can fully utilize a unified processing framework for multimodal data, and transfer existing models to the fields of text generation, 3D image generation, and video subtitle generation. With the continuous evolution of technology, we have ample evidence to infer that in the near future, it will surely achieve more significant breakthroughs.

References

[1] Gao, X. Y., Du, F., Song, L. J. (2024). Comparative Review of Text-to-Image Generation Techniques Based on Diffusion Modelsl. Computer Engineering and Applications. 60 (24).

[2] Jonathan, H., Ajay, J., Pieter, A. (2020). Denoising Diffusion Probabilistic Models. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems Article No.: 574, Pages 6840 – 6851.

[3] Li, L. Y., Tong, G. X., Zhao, Y. Z. et al. (2023). Survey of Text-to-Image Synthesis Based on Generative Adversarial Network. Electronic Sci. & Tech.

[4] Li, W. Y., Du, H. B., and Zhang, Q. (2025). Text generation image algorithm based on improved stable diffusion model and noise concatenation. Journal of Nanjing University of Information Science and Technology, 1-14.

[5] Liu, Z. R., Yin, F. Y., Xue, W. H. et al. (2023). A review of conditional image generation based on diffusion models. Journal of Zhejiang University (Science Edition) Volume 50, Issue 6.

[6] Mohamed, E., Omar, E., Somaya, A. et al. (2022). Image Generation: A Review. Article in Neural Processing Letters.

[7] Qiao, L. T., Zhang, J., Xu, D. Q. et al. (2019). MirrorGAN: Learning Text-to-image Generation by Redescription. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[8] Raffaella, B., Ruket, C., Desmond, E. et al. (2016). Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. Journal of Artificial Intelligence Research 55. 409-442

[9] Sibi, M. (2024). An Overview of Text to Visual Generation Using GAN. Indian Journal of Image Processing and Recognition (IJIPR)ISSN: 2582-8037 (Online), Volume-4 Issue-3

[10] Xu, Y. W., Chen, G. (2025). Text Matching Image Generation Model Based on Improved GAN Algorithm. Journal of Jilin University (Information Science Edition) Volume 43, Issue 2