Text to Image Generation: A Literature Review Focus on the Diffusion Model

Jingxi Zhou*

Beijing No. 80 High School, Beijing, 100102, China

Abstract. This paper reviews the progress in text-to-image generation, which enables the creation of images from textual descriptions. This technology holds promise across various fields, including creative arts, gaming, and healthcare. The main approaches in this area are Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models (DM). While GANs initially made significant advancements in realistic image generation, they faced issues with stability and diversity. VAEs introduced a probabilistic approach, allowing for diverse outputs but often at the cost of image quality. The development of DM, like Stable Diffusion, Imagen, and DALL-E 2, has addressed many limitations, producing high-quality, coherent images through iterative denoising. DM stands out for its stability and ability to generate detailed, semantically accurate images. This review explores the strengths and limitations of each approach, with an emphasis on the advantages of DM. It also discusses future directions, including improving efficiency, enhancing multimodal capabilities, and reducing data requirements to make these models more accessible and versatile for various applications.

1 Introduction

Text-to-image generation (TIG) is a technology with vast potential in advertising, gaming, education, and healthcare fields. Combining Natural Language Processing (NLP) with computer vision, enables the transformation of text descriptions into realistic images, opening up new possibilities for various applications.

Based on deep learning, there are three main streams of TIG, including the method based on Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion model (DM).

In 2014, GANs were introduced, marking the first significant advancement in this field. GANs introduced a fresh approach to picture synthesis. Image realism significantly improved as a result of this invention. Mode collapse, in which the generator creates images with much less diversity and crosses a convergence to produce unreliable images, was a problem for GANs. Furthermore, because the learning process is adversarial, training GANs were frequently unstable, producing inconsistent results [1]. Despite their widespread use, GANs' practical usefulness has been constrained by these issues [2].

* Corresponding author: Jayden080603@outlook.com

To overcome some of these issues, VAEs, introduced by Kingma offered an alternative generative approach by utilizing probabilistic modeling. VAEs introduced a framework that generated diverse outputs by encoding inputs into a latent space. However, this method often produces blurry images due to the nature of the reconstruction process, particularly when dealing with high-resolution or complex scenes [3, 4].

The introduction of DM marked a turning point in TIG, offering an approach that addressed many of the limitations seen in GANs and VAEs. DM operates by iteratively adding noise to data and then learning to reverse this process to generate new images. This denoising approach results in models that are more stable during training and capable of producing highly detailed, semantically aligned images. The iterative nature of DM allows them to avoid model collapse and produce outputs that are both diverse and of high quality [5, 6].

This review examines the evolution of TIG from GANs to DM, comparing their principles, strengths, and limitations. This research focuses on recent DM due to their ability to address the shortcomings of earlier approaches, such as mode collapse and blurry outputs, and their applications across a variety of fields. Finally, this paper explores future research directions in optimizing these models for greater computational efficiency and expanding their use to multimodal generation tasks

2 Key methods in TIG

2.1 GANs

GANs consist of two networks: a generator and a discriminator. The generator creates images from random noise, while the discriminator evaluates the authenticity of both the generated and real images. This adversarial training is represented by the following objective function.

$$V(G,D) = E_{x\sim P_{data}}[logD(x)] + E_{x\sim P_G}\left[\log\left(1 - D(x)\right)\right] \quad (1)$$

$$G^* = arg\min_G \max_D V(G,D) \quad (2)$$

The generator tries to minimize $log(1 - D(G(z))$, i.e. it wants the discriminator to think the generated samples $G(z)$ are real. In contrast, the discriminator ties to maximizing $log(x) + log(1 - D(G(z))$, i.e., it wants to correctly classify the real data x and generated samples $G(z)$. In the objective function, $E_{x\sim P_{data}}$represents samples drawn from the true data distribution P_{data}whereas $\left[\log\left(1 - D(x)\right)\right]$ aiming to make $G(z)$ appears as real as possible.

The process continues until a Nash equilibrium is reached, where the discriminator cannot distinguish between real and generated data better than random guessing.

Although GANs are remarkably good at high-quality outputs and unsupervised learning, they are often difficult to train and will easily collapse without carefully selected hyperparameters and regularization [7, 8].

2.2 VAEs

VAEs is a probabilistic generative model designed to learn latent space representations of data, particularly in TIG. The main mechanism involves both encoding and decoding processes. In the encoding process, data is input into a continuous, lower-dimensional latent space, whereas the decoding process then decodes it back to its original or transformed form.

The advantages of VAEs model include the introduced random variations during the decoding process, allowed by the probabilistic formulation, leading to the generation of diverse outputs. However, the smoothness of the latent space, while enabling diversity, often leads to well-documented challenges in producing sharp images, as the decoder struggles to recover high-frequency details from the latent variables [9, 10]. To address this problem, Ali Razavi et.al. introduced a multi-level hierarchical latent space, enabling VQ-VAE-2 to significantly enhance the models' capacities to represent both broad and fine-grained features, resulting in improved image reconstructions with finer details [9].

The hierarchical design of VQ-VAE-2, while beneficial for enhancing image quality, comes with the trade-off of higher computational demands. This includes greater memory usage and longer training times, particularly when generating high-resolution images.

2.3 DM in TIG

2.3.1 Principles of DM

DM creates new data by progressively adding noise to existing data and learning to reverse this process during denoising. This approach is based on a Markov chain, transforming random noise into structured images through noise reversal. The text input guides the denoising process, ensuring alignment between the generated images (GI) and the provided text. Unlike VAEs and GANs, DM follows a multi-step approach, enhancing stability, image quality, and text-to-image alignment [11].

Denoising Diffusion Probabilistic Models (DDPMs) are commonly used for text-to-image tasks in models like DALL-E 2, Imagen, and Stable Diffusion (SD), excelling in producing high-resolution, detailed images and outperforming earlier models that struggled with issues like mode collapse and blurry outputs.

2.3.2 Early foundations of DM

The development of DM began with Denoising Diffusion Probabilistic Models (DDPMs) by Ho, which introduced iterative denoising to generate high-fidelity images [12]. Unlike earlier generative models, DDPMs provided more control over output refinement through multiple iterations. However, early models operated in pixel space, leading to high computational costs, which limited their practical applications. Yang's survey [13] on DM highlighted these computational challenges, emphasizing the need for faster, high-quality solutions.

2.3.3 TransitioAAAn to Latent Space for efficiency

A major advancement in DM was the shift from pixel space to latent space, as seen in SD. Latent DM (LDMs) use pre-trained autoencoders to compress images into a latent representation, focusing on essential visual features, which enhances computational efficiency without compromising quality [14]. This improvement makes them more suitable for real-time image generation in industries like advertising, gaming, and design. Latent space diffusion reduces resource demands, allowing models to operate on consumer-grade hardware, and expanding applications beyond static images to areas like forecasting and simulation.

2.3.4 Classifier-Free Guidance (CFG) for Enhanced Control

CFG allows for more flexible control over the generation process, integrating guidance directly into the diffusion process without requiring separate classifiers. This method enables users to adjust the balance between text prompt fidelity and image diversity. For instance, Imagen leverages this technique to create photorealistic images closely aligned with complex prompts, useful in fields requiring highly specific visual outputs. The ability to fine-tune models for specific subjects, as explored in Dream Booth, allows for personalized content generation in fields like brand-specific designs or personalized artwork [15].

2.3.5 Key DM

Imagen Strengths: High-Resolution Outputs: Imagen consistently delivers high-quality images that are detailed and photorealistic visuals.

Photorealism: Imagen excels in generating photorealistic images. By leveraging large-scale language models like T5, it effectively captures complex text prompts, enabling it to generate intricate scenes with accurate object placement and relationships [16].

CFG: Imagen's use of CFG allows fine-tuning of image outputs, balancing fidelity and diversity, which is particularly useful for creative projects where varying degrees of control over image generation are needed.

Weaknesses: Computational Demand: Despite its strengths, Imagen's high-quality output comes at the cost of increased computational requirements. The model's need for extensive resources can make it less accessible for smaller-scale applications or real-time use.

SD Strengths: Efficient Latent Space Diffusion: SD's use of latent diffusion dramatically reduces computational costs by operating in a lower-dimensional latent space. This makes the model more efficient without sacrificing image quality, enabling high-resolution images to be generated with fewer resources [17].

Accessibility: SD is an open-source model, widely available to developers and businesses, which has fueled its adoption across a range of industries, from creative arts to marketing and gaming. It can also run on consumer-grade hardware, making it one of the most versatile models for both commercial and personal use.

Comparative Summary: Imagen stands out for its photorealism and fine control, making it ideal for industries that require detailed, high-quality visuals and complex scene generation. However, its high computational demand limits its accessibility for real-time or lightweight applications.

SD offers a balanced solution between computational efficiency and quality. It's more accessible due to its open-source nature and latent space optimization, which allows it to run on less powerful hardware. However, its lower resolution and the complexity of tuning may be limiting factors for some applications.

DALL-E 2 is a powerhouse in creative generation, excelling in imaginative and diverse outputs. However, its slower inference times and limited control over fine-grained details make it less suitable for precision-demanding tasks compared to Imagen or SD.

2.3.6 Technical Challenges in Text-to-Image DM

Despite significant progress in text-to-image DM, technical challenges persist:

Computational Efficiency: Although latent space diffusion has dramatically improved the efficiency of image generation, producing complex, high-resolution images remains computationally demanding. Pixel-space models like Imagen require extensive resources, particularly for generating photorealistic images. Balancing computational efficiency with

image quality remains an ongoing challenge, especially for real-time applications where latency is a concern [14].

Training Stability: Training stability continues to be a significant issue in DM, particularly when dealing with large and diverse datasets. The paper Reflected DM for Generative Tasks [18] addresses these challenges by proposing solutions that improve model stability during training, ensuring that DM generalizes well across different domains. This is particularly relevant for fields like virtual reality or scientific visualization, where generating consistent, high-quality images is crucial.

Fine-Grained Control Over Image Attributes: While models like SD and DALL-E 2 offer flexibility in TIG, achieving fine-grained control over specific features such as texture, lighting, and perspective remains challenging. Ongoing research aims to enhance the precision of these controls to allow users more granular manipulation of the output images, as highlighted in Artificial Intelligence-Generated Content with DM and Dream Booth [15, 19].

3 Experiment

3.1. Dataset introduction

The experiments focus on the Common Objects in Context dataset (COCO), which is an accepted standard for text-to-image tasks. This dataset provides a broad spectrum of objects, scenes, and relationships, containing over 330,000 images and 80 object categories.

3.2. Algorithm performance comparison

Below is a detailed comparison of popular models across several key performance metrics. It examines models using the FréChet Inception Distance (FID), which measures the similarity between GI and real images; the CLIP score, which assesses how well GI aligns with the given text prompts; the Inception Score (IS), which assesses GI diversity; and Perceptual Image Quality (PIQ), which assesses image realism based on human perception. These criteria are broadly acknowledged within the research community [20].

Table 1. The comparison of different generative models

Model	FID(COCO)	CLIP Score	Inception Score (IS)	Perceptual Image Quality (PIQ)	Resolution	Inference Time
Attn GAN	35.49	N/A	12.7	0.63	256*256	15s
DALL-E 2	10.39	88.0	45.2	0.89	1024*1024	10s
GLIDE	12.24	90.5	43.5	0.87	1024*1024	12s
Imagen	7.27	95.0	50.8	0.92	1024*1024	9s
VQ-VAE 2	15.78	N/A	28.6	0.68	256*256	30s
Beta-VAE	27.45	N/A	21.3	0.57	256*256	25s
SD	8.15	91.7	47.9	0.90	786*786	25s

3.3. Experimental results

In Table 1, DM like Imagen outperforms traditional GAN-based models in terms of both fidelity and diversity. DM generates images in a stepwise manner, allowing for more detailed and refined outputs. In contrast, models like VAEs and GANs often struggle with issues like blurred outputs (VAEs) or mode collapse (GANs), especially when dealing with high-resolution image generation tasks.

Furthermore, while Imagen and GLIDE use cascaded DM to upscale low-resolution images to high-resolution outputs (1024x1024), GAN models like AttnGAN operate at much lower resolutions (256x256), which limits their applicability in scenarios requiring high visual detail.

These results highlight the continued evolution of text-to-image models, where DM now represent the state of the art due to their superior performance in both image quality and efficiency.

4 Future directions

4.1 Improving computational efficiency in TIG

As the diffusion model became state-of-the-art. The efficiency of text-to-image DM is a key area of development, especially as the demand for real-time applications increases. Recent progress in distillation methods has enabled models like Swift Brush to generate high-quality images with only a single step, significantly reducing the computational load while maintaining the fidelity of the GI. Swift Brush, for instance, achieves image quality comparable to SD using far fewer sampling steps, making it highly suitable for interactive applications like real-time graphic design or creative tools [21].

Similarly, Mobile Diffusion provides another critical innovation by allowing text-to-image models to run efficiently on mobile devices. By employing extensive architectural optimizations and knowledge distillation, Mobile Diffusion has achieved sub-second inference times for generating 512x512 images, a breakthrough for integrating advanced generative models into everyday consumer technology. This direction points towards a future where text-to-image DM are not just faster but also more accessible across a variety of platforms, from desktop computers to mobile phones.

4.2 Expanding multimodal capabilities

Although current DM is largely focused on 2D image generation from text, future directions are steering towards more complex outputs, including 3D objects and even video. Some models are starting to bridge the gap between 2D TIG and 3D synthesis, using shared diffusion processes. Research into this area promises significant advancements in industries like virtual reality, gaming, and immersive simulations, where generating detailed 3D environments from textual descriptions will open new possibilities for creative work and user engagement. This shift will likely involve integrating additional sensory modalities such as audio, further enriching the user's experience in creative and interactive media.

4.3 Reducing data requirements for training text-to-image models

One of the ongoing challenges for DM, especially those applied to text-to-image tasks, is their dependency on vast training datasets. Recent research on model architectural compression and knowledge distillation has focused on reducing these data requirements. For

instance, models like Block-Knowledge Distilled SDMs (BK-SDMs) have successfully reduced their parameter count while achieving competitive results on benchmarks with a significantly smaller dataset. These techniques enable more lightweight models that can be trained with fewer resources, making TIG more feasible for specialized domains like healthcare, where data availability is limited.

5 Conclusion

In conclusion, TIG using DM has demonstrated considerable advancements, surpassing previous approaches like GANs and VAEs in both image fidelity and semantic alignment. DM have proven particularly successful in maintaining image diversity and avoiding common pitfalls like mode collapse, which plagued earlier methods. This stability and performance improvement is largely attributed to the iterative denoising processes inherent to DM, allowing for high-quality image outputs even from complex textual prompts.

However, despite these advancements, challenges remain, particularly in the realm of computational efficiency. As these models continue to gain traction across industries—spanning from entertainment to healthcare, the need for faster and more resource-efficient models becomes critical. The recent shift towards latent space diffusion, as seen in models like SD, represents a major step in this direction by reducing computational demands without sacrificing image quality. Similarly, innovations like Swift Brush's single-step generation highlight the potential for real-time applications, which could revolutionize fields such as design and interactive media.

Looking ahead, the integration of multimodal capabilities presents a promising frontier. The ability to generate 3D objects and video from text, while still in its early stages, could vastly expand the applications of DM in immersive environments such as virtual and augmented reality. Moreover, enhancing these models to incorporate additional sensory inputs, like audio, could further transform user experiences, offering more interactive and rich content creation tools.

Finally, reducing the data requirements for training DM is an important focus for future research. The development of techniques such as architectural compression and distillation can make these models more accessible in domains where data is scarce or expensive to acquire, such as medical imaging. These advancements will not only broaden the applicability of DM but also democratize access to powerful generative technologies across various fields.

In summary, while DM has already pushed the boundaries of TIG, ongoing improvements in efficiency, multimodal capabilities, and data usage will likely solidify their role as a cornerstone of future AI-driven creativity and industrial applications.

References

1. I. Goodfellow, et al., Generative Adversarial Nets, NeurIPS (2014).
2. A. Radford, L. Metz, S. Chintala, Unsupervised Representation Learning with Deep Convolutional GANs, ICLR (2015).
3. D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes, arXiv preprint arXiv:1312.6114 (2013).
4. A. Razavi, et al., Generating Diverse High-Fidelity Images with VQ-VAE-2, Advances in Neural Information Processing Systems (NeurIPS) (2019).
5. C. Saharia, et al., Imagen: Photorealistic Text-to-Image DM, Google AI Research (2022).

6. A. Ramesh, et al., DALL·E 2: A Diffusion Model for Text-to-image generation, OpenAI (2022).
7. P. Dhariwal, A. Nichol, DM Beat GANs on Image Synthesis, arXiv preprint arXiv:2105.05233 (2021).
8. I. Goodfellow, et al., Generative Adversarial Nets, NeurIPS (2014).
9. C. Doersch, Tutorial on Variational Autoencoders, arXiv:1606.05908 (2016).
10. D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes, arXiv:1312.6114 (2014).
11. S. Ge, T. Park, J. Y. Zhu, J. B. Huang, Expressive Text-to-image generation with Rich Text, Proceedings of the IEEE/CVF International Conference on Computer Vision, 7545-7556 (2023).
12. J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in Neural Information Processing Systems **33**, 6840-6851 (2020).
13. L. Yang, Z. Zhang, Y. Song, et al., DM: A comprehensive survey of methods and applications, ACM Computing Surveys **56**(4), 1-39 (2023).
14. R. Rombach, A. Blattmann, D. Lorenz, et al., High-resolution image synthesis with latent DM, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684-10695 (2022).
15. N. Ruiz, Y. Li, V. Jampani, et al., Dreambooth: Fine tuning text-to-image DM for subject-driven generation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 22500-22510 (2023).
16. C. Saharia, W. Chan, S. Saxena, et al., Photorealistic text-to-image DM with deep language understanding, Advances in Neural Information Processing Systems **35**, 36479-36494 (2022).
17. J. Ho, T. Salimans, Classifier-free diffusion guidance, arXiv preprint arXiv:2207.12598 (2022).
18. A. Lou, S. Ermon, Reflected DM, International Conference on Machine Learning, PMLR, 22675-22701 (2023).
19. X. Wang, Z. He, X. Peng, Artificial-Intelligence-Generated Content with DM: A Literature Review, Mathematics **12**(7), 977 (2024).
20. C. Saharia, W. Chan, S. Saxena, et al., Photorealistic text-to-image DM with deep language understanding, Advances in Neural Information Processing Systems **35**, 36479-36494 (2022).
21. T. H. Nguyen, A. Tran, Swiftbrush: One-step text-to-image diffusion model with variational score distillation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7807-7816 (2024).