Text-to-image generation: From GANs and Diffusion Models to LLM-Augmented Paradigms

Ran Tao

{2410240620@henu.edu.cn}

International School of Technology, Henan University, Zhengzhou, Henan, China

Abstract. As one of the core cutting-edge fields of AI-generated content, text-to-image generation has brought about dramatic changes to traditional content creation fields. This paper divides the development of text-to-image generation into three stages: Generative Adversarial Networks (GAN), Diffusion Models and the exploration of Large Language Model (LLM) Augmented Paradigms. Based on the current representative technological evolution paths, LLM-enhanced text-to-image generation methods are further summarized into three main paradigms: LLM-based Text Encoder Enhancement, LLM-based Caption & Prompt Refinement, LLM-grounded Layout Planning. Although current LLM enhancement methods still have certain limitations, they have shown great potential in semantic understanding and generative control and are expected to further promote the development of text-to-image technology. This research dissects the progression of text to image generation and outlines three stages for LLM enhancement. It not only offers a clear theoretical framework and practical direction for further research in the domain, but also enriches the understanding of the challenges and opportunities related to the semantic understanding and controllability of generated content, possessing significant academic value and application significance.

Keywords: Text-to-image generation, GAN, Diffusion, LLM

1 Introduction

In this period of incredibly fast AI development, the Artificial Intelligence Generation Content (AIGC) has obviously become one of the pressing concerns. With the development of Computer Vision (CV) and Natural Language Processing (NLP), text to image generation has evolved at an exponential rate and is currently among the most significant fields in AIGC. Text-to-image generation is the process of generating images, which are of high quality, based on natural language text descriptions. Its text-to-image generation, which is very powerful, has helped overcome the barrier to entry into the creative process, as well as increasing productivity, by creating a significant influence in such areas as creative expressiveness, design, and multimedia content creation [1].

During the course of the years of development, the field of text-to-image generation has transformed developing from the simple types of images to more complex and realistic ones.

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365347

One of the first attempts to apply attention-based mechanisms to the literature to produce images was made as early as 2015 with Mansimov et al. being the first researchers to attempt the use of this method within this field [2]. Generative Adversarial Networks (GANs) were introduced by Reed et al. as the backbone generative model in 2016 [3]. As a further extension, Attentional Generative Adversarial Networks (AttnGAN) were introduced by Xu et al. as GANs with attention mechanisms incorporated into the framework to provide high-quality image generation with a high degree of granularity and fidelity [4]. Later, Dhariwal and Nichol both proved that diffusion models are entirely superior to GANs in terms of the quality of images generated [5]. More to the point, Rombach et al. using Latent Diffusion Models (LDMs), made breakthroughs in high-resolution generation and the ability to train on a single-gpu thus again making text to image generation massively cheaper in terms of the cost of computing [6]. Nevertheless, even current models have severe semantic lapses facing the problem of complex space logics and reading long texts. As a result, Saharia et al. presented Imagen in 2022 that showed that text-to-image generation models could be integrated with Transformers [7]. At this point, the use of Large Language Models to empower text-to-image generation has become a top frontier movement.

Despite the fact that many articles have already explained the evolution of this field by chronological reviews, which is simply a list of models and the time they were developed, this paper will not be chronological but will follow the logic of a syllogism - Problems Threats solutions: What troubles were solved by previous models- What do the future generations fail to solve.

2 GAN-based text-to-image generation

Early text-to-image generation algorithms (including Conditional GANs) were mainly based on the compression of the whole text description into a single global sentence vector as conditioning. This method coded the fine-grained information in the text which was not easy to get the specifics of the description of various objects through the generative model.

To deal with the above pain point, AttnGAN proposes the Attentional Generative Network and Deep Attentional Multimodal Similarity Model (DAMSM), which enhances the truthfulness of image production by the sentence instead of the word [4]. AttnGAN uses the Attentional Generative Network to allow it to take a multistage generation approach. To start with, a low-resolution bluring is developed on the basis of global features. Second, the detail is generated at a greater resolution by applying an attention mechanism. The model automatically retracts and concentrates on the word which is most common in the present image sub-region. This facilitates the model to give fined synthesis on various sections in the image. In order to be consistent between the text descriptions and the created fine-grained images, AttnGAN adds DAMSM loss function to compute the matching degree between image sub-regions and words, hence giving stronger fine-grained supervision indications when training.

Although the AttnGAN has achieved a great deal in terms of fine-grained control, the nature of the GAN architecture predisposes it to certain issues, including the lack of diversity of representations and the inability to train. That compels the research community to resort to other new generative paradigms.

3. DM-based text-to-image generation

Earlier Diffusion operations were operated directly in the high-dimensional Pixel Space, which inferred images at extremely high costs, with carbon footprint becoming so detrimental that it severely hindered their usage deployment.

In order to solve the above problems, LDMs suggested Perceptual Image Compression & Latent Space and Cross-Attention to preserve high fidelity and diversity and manage time and cost [6]. In LDMs, in contrast to direct processing pixels, first an autoencoder that was trained to compress the image into a low-dimensional latent space is used. Diffusion and the denoising are done in this low-dimensional space alone and eventually the image is re-assembled with the help of a decoder and thereby time and cost are controlled. More so, by integrating a Cross-Attention layer into the U-Net backbone network, LDMs have the capability to arbitrarily add textual conditions, which can be trained with exceptionally little computing power (can be handled by a single consumer-grade graphics card) while simultaneously achieving a very high level of fidelity and diversity.

Stable Diffusion alleviated the issues of quality and efficiency in generation, but it left a gap in the important semantic aspects. The models have difficulties with interpreting the complex logical relationships or a long text description because they are based on relatively simple text encoders (including CLIP) [8]. This is what brought us to the present day large language model systems era.

4. LLM-based text-to-image generation

4.1 LLM-based text encoder enhancement

The CLIP operated as a text encoder used to power all the previous mainstream models (including LDMs and DALL-E 2) that were also trained on image-text pairs. In turn, CLIP does not have the unique nature of text and variety of images; this leads to poor quality image generation in the presence of complicated and logical relationships as well as lengthy descriptions of texts [8].

To overcome the failure of more traditional text encoders, like CLIP, to understand more complicated prompts and logic, Imagen directly uses a pre-trained LLM, like T5, to substitute the text encoder [7]. The scale of LLM is usually many times greater than that of the text encoders of the modern picture-text models; therefore, they are presented with a very wide and large variety of text data. This will allow Imagen to get significant gains in the quality of generated images as well as text-image alignment.

Even though the introduction of LLMs has remarkably improved language comprehension systems, the magnitude of such models, even exceeding that of the generative U-Net itself, makes them extremely expensive computationally and incurs a massive inference time.

4.2 LLM-based caption & prompt refinement

Because of the presence of low-quality text-image pairs in the datasets that are used to learn the old models, the models often have issues with the concept of prompt following in image

generation, that is, they do not pay enough attention to the exact words, word sequence, or semantic content of a certain caption [9]. Moreover, average users usually just give short and general prompts that do not include particular details; as a result, the images that are generated do not match the expectations the users have.

In order to resolve such problems, the Betker team suggested a new method known as caption improvement and used it on the DALL-E 3 model [9]. They first trained a powerful image captioning model and also used it on their dataset to create more detailed descriptions; they then trained a text-to-image generation model on their dataset with enhanced descriptions. Moreover, Hao et al. introduced the prompt optimization loop based on the model, where supervised fine-tuning of pre-trained models is done to perform systematic adjustment of the user intentions with different model-preferred prompts [10].

Despite making enormous progress in ensuring the quality of generated images by integrating the idea of LLMs to maximize the captions and prompts, there is still a problem with the location of the objects and their awareness of space. Moreover, when optimizing captions and prompts, the LLMs are prone to causing hallucinations, which are the creation of non-present details about the important aspects of the picture.

4.3 LLM-grounded layout planning

Conventional text encoders (e.g., CLIP) have a hard time understanding complicated logic and encoding quantitative relationships in the best way, so they do not produce the best quality images. The earlier solutions included training using more extensive datasets; however, this is time and cost-intensive.

Due to the above, Lian et al. suggested LLM-grounded Diffusion (LMD)- a two-stage, training-free, and generation pipeline [11]. In stage one, they proposed an LLM as a grounded text layout, which is inspired by a prompt of what image is needed, which is a scene layout represented as captioned bounding boxes, and thereby provides a better interpretation of the prompt by the text-to-image generation models. At the second stage, the image is created under the direction of a new controller with the help of a diffusion model and in accordance with the layout generated during the first stage. The method allows specific objects to be controlled within specific regions, as opposed to the previous methods, where the semantic controls were applied in large regions of space.

Despite the fact that the process of LMD, which includes the two stages, helps to enhance logical understanding, the generation process, split into two stages, is bound to introduce cascading errors. As soon as a mistake is made at the first stage, the latter will fail as well upon the second stage, which will not have any corrective power, and it will ultimately result in the irreversible structural damage of the image.

5. Conclusion

The paper undertakes a review of the developments in the area of text-to-image generation; the review logic is based on the principles of "Pain Points-Innovations-Limitations". The GAN era was the first production of high-quality text-to-image generation, but the natural weaknesses that made it hard to train the model ultimately sent the field to the Diffusion paradigm.

Although Diffusion was able to address problems related to training stability, it used text encoders; therefore, its performance was usually poor in the presence of complex prompts. This has led to the field entering an exploration stage that is aimed at putting the power of LLM into text-to-image generation. This exploration mainly includes three large paradigms as follows: LLM-based Text Encoder Enhancement, LLM-based Caption and Prompt Refinement, and LLM-grounded Layout Planning. The three paradigms are mutually complementary and work in parallel, where each one of them attempts to address the weaknesses of the traditional text encoders by focusing on different aspects. At the same time, a certain trend has also been observed toward the fusing of these three methods; next-generation generative systems will cease to just be the simple concatenation of a set of independent modules but will become integrated into a multimodal architecture formed naturally. The current landscape has already seen the mature products, which include banana pro nano. In the years to come, LLMs will be fully able to capitalize on their remarkable text understanding abilities and address the difficulties linked to comprehending intricate logical prompts and drive this direction to new heights.

References

[1] Bie, F., Yang, Y., Zhou, Z., Ghanem, A., Zhang, M., Yao, Z., et al.: Renaissance: A survey into ai text-to-image generation in the era of large model. IEEE Transactions on Pattern Analysis and Machine Intelligence. 47(3), pp. 2212-2231 (2024)

[2] Mansimov, E., Parisotto, E., Ba, J.L., Salakhutdinov, R.: Generating images from captions with attention. arXiv preprint arXiv:1511.02793 (2015)

[3] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning, pp. 1060-1069 (2016)

[4] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1316-1324 (2018)

[5] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems. 34, pp. 8780-8794 (2021)

[6] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684-10695 (2022)

[7] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems. 35, pp. 36479-36494 (2022)

[8] Wang, W., Sun, Q., Zhang, F., Tang, Y., Liu, J., Wang, X.: Diffusion feedback helps clip see better. arXiv preprint arXiv:2407.20171 (2024)

[9] Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., et al.: Improving image generation with better captions. Computer Science. 2(3), pp. 8 (2023). https://cdn.openai.com/papers/dall-e-3.pdf

[10] Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. Advances in Neural Information Processing Systems. 36, pp. 66923-66939 (2023)

[11] Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint arXiv:2305.13655 (2023)