Proceedings of CONF-MLA 2024 Workshop: Neural Computing and Applications
DOI: 10.54254/2755-2721/105/2024TJ0053

Enhancing Text-to-Image Generation: Integrating CLIP and Diffusion Models for Improved Visual Accuracy and Semantic Consistency

Ziyang Wang

College of Computer Science and Technology, Jilin University, Jilin, China

wangzy2121@mails.jlu.edu.cn

Abstract. Text-to-Image (T2I) generation focuses on producing images that precisely match given textual descriptions by combining techniques from computer vision and natural language processing (NLP). Existing studies have shown an innovative approach to enhance T2I generation by integrating Contrastive Language-Image Pretraining (CLIP) embeddings with a Diffusion Model (DM). The method involves initially extracting rich and meaningful text embeddings using CLIP, which are then transformed into corresponding images. These images are progressively refined through an iterative denoising process enabled by diffusion models. Comprehensive experiments conducted on the MS-COCO dataset validate the proposed method, demonstrating significant improvements in image fidelity and the alignment between text and images. When compared to traditional models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), which often struggle with maintaining both visual quality and semantic accuracy, this hybrid model shows superior performance. Future research could explore optimizing hybrid models further and applying T2I technology to specialized fields, such as medical imaging and scientific visualization, expanding its potential use cases.

Keywords: Text-to-Image, Diffusion Models, CLIP, Generative Adversarial Networks (GANs).

1. Introduction

The ability to generate high-quality images from descriptions, referred to as Text-to-Image (T2I) generation, has gained increasing attention within both academic and industrial contexts. This capability bridges computer vision and natural language processing (NLP), allowing models to transform textual inputs into detailed visual outputs. T2I generation has diverse applications, ranging from content creation and augmented reality to AI-assisted design. Moreover, it serves as an important benchmark for multimodal learning, where models are required to understand and generate content across multiple domains, such as text, image, and audio [1, 2].

Advancements in generative modeling have significantly propelled progress in T2I generation. Generative Adversarial Networks (GANs) were among the first models to successfully produce high-quality images from text, leveraging a competitive training setup involving a generator and a discriminator [3]. Despite their impressive performance, GANs suffer from limitations, such as mode collapse. At this time the generator produces a narrow range of images, during training, which reduces image diversity and robustness [4]. In response, alternative models such as Variational Autoencoders

DOI: 10.54254/2755-2721/105/2024TJ0053

(VAEs) and Diffusion Model (DMs) have been explored. VAEs offer a probabilistic framework for stable training and interpretable latent spaces, though they typically generate lower-quality images compared to GANs [5].

DMs, in contrast, have acquired attention for their unique ability to generate multiple, high-quality images by refining random noise through iterative denoising. Unlike GANs, which generate images in a single pass, DMs operate through multiple steps, allowing for more controlled generation and greater stability during training [6, 7]. This refinement process strikes a balance between fidelity and diversity, making DMs an attractive choice for T2I generation.

One of the key breakthroughs in T2I generation has been the integration of multimodal pretraining techniques, particularly Contrastive Language-Image Pretraining (CLIP) embeddings. CLIP is trained on large-scale paired datasets of text and images, capturing intricate semantic relationships between language and visual content [8]. By combining CLIP embeddings with generative models, researchers have created systems that not only generate photorealistic images but also maintain strong semantic alignment with the input text. These models have given impressive zero-shot capabilities across many tasks, allowing for the generation of images based on novel prompts.

However, T2I generation remains a complex task due to the inherent challenges in modeling both textual and visual content simultaneously. Textual descriptions can vary in specificity and detail, while visual perception is inherently subjective, adding layers of complexity to generating both accurate and aesthetically pleasing images. Furthermore, the computational costs of training advanced models like GANs, VAEs, and DMs are significant, requiring extensive resources for optimal performance.

Ongoing research in T2I generation explores new model architectures, training strategies, and multimodal integration methods to overcome these challenges. Recent studies propose composite models that combine the strengths of GANs, VAEs, and DMs to generate higher-quality images with improved diversity and semantic coherence [9]. There is also a growing interest in enhancing the interpretability and controllability of T2I models, possibly through user feedback or the development of more sophisticated text encodings.

This paper reviews the current state of T2I generation, focusing on the evolution of generative models and recent advancements in multimodal learning. It highlights the key challenges facing the field and proposes future research directions. Specifically, it emphasizes the need for more efficient models that can generate high-quality images at lower computational costs, as well as the exploration of specialized applications where T2I models can add significant value, such as in medical imaging and scientific visualization.

2. Methodology

This section details the methodology used to enhance T2I generation. The MS-COCO dataset was employed for training and evaluation, and an innovative approach integrating CLIP embeddings with Diffusion Models was proposed.

2.1. Dataset description and preprocessing

The MS-COCO dataset was chosen for this study because it is known for its large scale, diversity, and high-quality images that are widely used in computer vision [9]. The dataset has over 330,000 images, each accompanied by five descriptive captions, making it ideal for tasks that require a high degree of consistency between textual captions and visual content. The detailed captions have the model to learn the complex relationship between language and image generation.

To prepare the dataset for training, this paper implemented various preprocessing steps. First, the images were resized for consistency and the pixel values were normalized. Textual descriptions were tokenized, converted to lowercase, and stop words and non-alphabetic characters were removed to reduce noise and improve model generalization, thereby increasing training efficiency. Furthermore, this study employs enhancement techniques like random cropping, horizontal flipping, and color dithering to expand the dataset's diversity. These methods help to reduce the likelihood of overfitting and enhance the model's overall robustness, ensuring more reliable performance during training.

DOI: 10.54254/2755-2721/105/2024TJ0053

2.2. Proposed approach

The aim of this research (show in Figure 1) at first is to strengthen the visual and semantic accuracy of T2I generation by leveraging detailed textual prompts within a diffusion-based generative framework. This study addresses the inherent challenges in generating images that align accurately with given text by integrating CLIP embeddings with DMs. The goal is to develop a model capable of producing high-quality images that are both photorealistic and semantically consistent with the input text. By combining the strengths of CLIP, which captures deep semantic relationships between language and images, with the iterative refinement process of DMs, this approach presents a important advancement in the field of T2I generation.

The research framework consists of four key sections designed to provide a comprehensive understanding of the methodology. The initial part provides a detailed overview of the T2I task, highlighting the technological developments that have driven advances in generative modelling. Well-known models such as GANs, VAEs and DMs are examined, highlighting their strengths and limitations in generating high-fidelity and varied images. GANs are widely acclaimed for generating clear, high-resolution images, but they often face problems such as pattern collapse, resulting in a limited variety of output images. In contrast, VAEs offer more stable training and an interpretable potential space, but they tend to produce lower quality images compared to GANs. Diffusion models have received increasing attention for their ability to iteratively improve noise into high-quality images, striking a balance between image fidelity and diversity that was difficult to achieve with earlier models.

The second component of the research delves deeper into the core principles of Diffusion Models, explaining the reverse diffusion process used to generate images. Diffusion Models work by gradually denoising a random noise sample over several iterations, progressively improving image quality with each step. This iterative approach considers more stable image generation compared to GANs, which generate images in a single pass and often suffer from instability during training. Diffusion Models also offer the advantage of producing fewer artifacts and generating more diverse images, making them particularly well-suited for tasks like T2I generation that require both visual accuracy and variety. The structured denoising process inherent to DMs enables finer control over the final output, which is crucial for ensuring that the given images are realistic and closely aligned with the input text.

The third component of the research focuses on the integration of CLIP embeddings with Diffusion Models, demonstrating how this combination can significantly enhance the performance of T2I generation. CLIP, a model trained on large-scale datasets of paired text and images, excels at capturing rich semantic relationships between language and visual content. By using CLIP to extract detailed text embeddings, the model can more effectively translate the semantic meaning of the input text into corresponding visual elements. This integration enables the system to generate images that not only exhibit photorealism but also maintain strong semantic consistency with the provided textual descriptions. The pretraining process of CLIP allows it to capture fine-grained semantic relationships between words and visual elements, making it highly effective for generating images that reflect the nuances of the input text. For example, a prompt describing "a red apple on a wooden table" would result in an image that accurately represents both the color and the context, thanks to the detailed embeddings provided by CLIP.

The fourth and final component of the research showcases the experimental results of the proposed approach, with a specific focus on its performance on the widely used MS-COCO dataset. The MS-COCO dataset, which contains over 330,000 images paired with multiple descriptive captions, is a standard benchmark for evaluating T2I generation models. A large number of experiments were made to compare the performance of the planned diffusion-based model with traditional models like GANs and VAEs. The results indicate that the integration of CLIP embeddings with Diffusion Models leads to important improvements in image quality, diversity, and semantic consistency. Specifically, the diffusion-based model achieved a lower Fréchet Inception Distance (FID) score, indicating superior image fidelity, and a higher Inception Score (IS), which reflects better image diversity. The model was able to generate images that were not only visually appealing but also closely aligned with the semantic content of the input text, outperforming GANs and VAEs in this regard.

DOI: 10.54254/2755-2721/105/2024TJ0053

Additionally, the experiments highlighted the advantages of DMs in mitigating common issues combined with GANs, like mode collapse, which can limit the variety of generated images. The iterative refinement process of DMs ensures that the generated images are more diverse and less prone to repeating patterns, a significant improvement over the often-narrow range of outputs produced by GANs. The diffusion model's ability to progressively improve image quality also allows for finer control over the generation process, making it possible to achieve a more precise cooperation between the text description and the generated image.

In conclusion, this research presents a powerful new approach to T2I generation by combining the strengths of CLIP embeddings and DMs. By leveraging the rich semantic understanding provided by CLIP and the iterative refinement process of DMs, the proposed model significantly improves both the visual and semantic accuracy of generated images. Extensive testing on the MS-COCO dataset demonstrates that this hybrid model outperforms traditional approaches, such as GANs and VAEs, particularly in terms of generating high-quality, semantically consistent images. Looking forward, future research could explore the development of even more efficient hybrid models, as well as the application of T2I technology in specialized domains such as medical imaging and scientific visualization, where both visual accuracy and semantic alignment are crucial. This study opens new possibilities for the future of T2I generation, providing a solid foundation for further advancements in this rapidly evolving field.

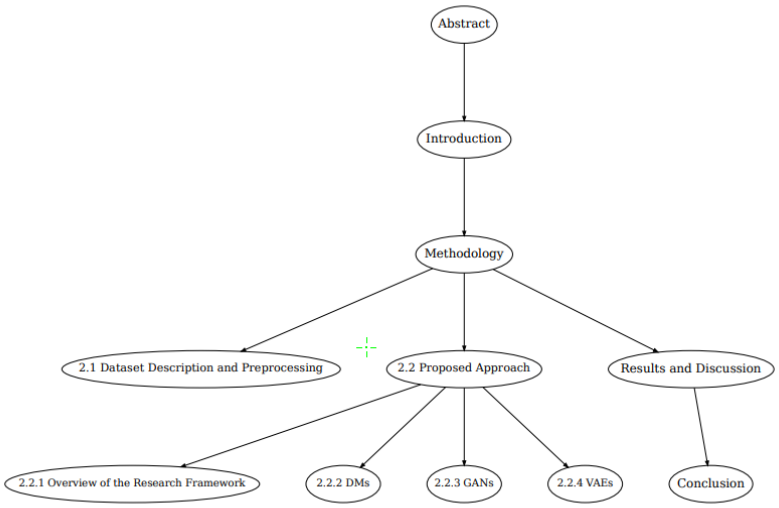

Figure 1. Structure of the research.

2.2.1. Diffusion models. DMs iteratively refine noisy images into high-quality visuals through a process of denoising. Unlike GANs, which generate images in a single step, DMs improve image quality step by step, offering greater control over the final output. A major strength of DMs is their ability to give nice images with fewer artifacts, as they avoid the adversarial setup found in GANs, which often leads to instability during training.

At the core of DMs is the modeling of data distributions as an opposite diffusion process. Originated from random Gaussian noise, an image is progressively denoised, with each step bringing it closer to the target output. The model is trained to predict the noise at each stage, effectively studying to contract the diffusion process and reconstruct high-quality images. This method has proven highly effective for high-resolution image generation, particularly for T2I tasks that require detailed and semantically consistent outputs [6, 10].

2.2.2. Generative adversarial networks. GANs are powerful models that produce high-resolution images using adversarial training. These models consist of two main parts: a generator that combines

DOI: 10.54254/2755-2721/105/2024TJ0053

images from random noise and a discriminator that evaluates how realistic the generated images are. Although effective, GANs often face challenges such as where the generator produces a limited range of images, and instability during training, which arises from the adversarial setup. To address these issues, models like StyleGAN and BigGAN have been developed, incorporating structural improvements to enhance training stability and improve the overall quality of the generated images.

2.2.3. Variational autoencoders. VAEs provide a probabilistic approach to generating images by encoding data into a potential space, from which the decoder reconstructs the images. VAEs ensure stable training and offer interpretable latent spaces, but they generally produce lower-quality images than GANs. The trade-off between latent space complexity and image fidelity remains a key limitation of VAEs, although they are particularly useful in applications requiring interpretable latent spaces and efficient image generation, such as medical imaging.

3. Result and Discussion

3.1. Results

As illustrated in Figure 2, the performance of various T2I models, including GANs, VAEs, and DMs, was assessed on the MS-COCO dataset using standard evaluation metrics, namely the FID and IS. These metrics provided a comparative analysis of the models' effectiveness. The DMs achieved a lower FID score of 14.9, indicating superior image fidelity compared to GANs (23.1) and VAEs (32.4). Additionally, the IS for DMs was higher at 4.2, suggesting better image diversity and quality over GANs (3.5) and VAEs (2.9) [10].

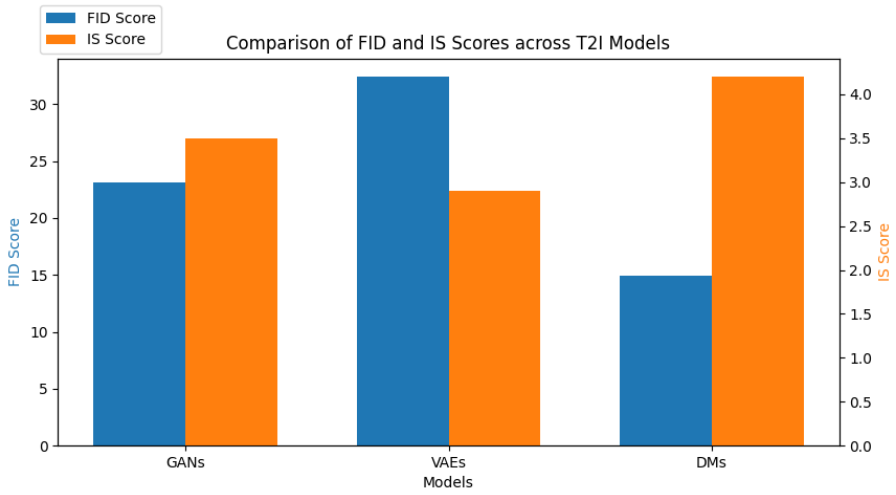

Figure 2. Comparison of FID and IS Scores across T2I models.

Table 1 summarizes the computational efficiency of each model, including training time and memory usage. Although DMs provide superior image quality, they require longer training times and higher memory usage due to their iterative denoising process. In contrast, GANs, while faster, suffer from issues like mode collapse, and VAEs offer moderate performance in both computational efficiency and image quality [11].

Table 1. Computational efficiency of each model.

Model	FID Score	Inception Score	Training Time (hours)	Memory Usage (GB)
GANs	23.1	3.5	15	8
VAEs	32.4	2.9	12	6
DMs	14.9	4.2	20	12

DOI: 10.54254/2755-2721/105/2024TJ0053

Table 1 summarizes the computational efficiency of each model, including training time and memory usage. Although DMs provide superior image quality, they require longer training times and higher memory usage due to their iterative denoising process. In contrast, GANs, while faster, suffer from issues like mode collapse, and VAEs offer moderate performance in both computational efficiency and image quality [11].

These results highlight the trade-offs between image quality and computational efficiency. While DMs offer the best performance in terms of image fidelity and diversity, they come at the cost of longer training times and increased memory usage. GANs, though faster, tend to collapse into generating repetitive images, while VAEs, although more computationally efficient, yield lower-quality images.

3.2. Discussion

Each T2I generation model evaluated in this study demonstrates unique strengths and weaknesses, largely driven by their underlying architecture. GANs excel at generating high-quality images in a relatively short time due to their adversarial setup, but they are prone to instability and mode collapse, which limits output diversity [2, 4]. VAEs provide a more stable training process, but their trade-off between latent space complexity and image fidelity leads to lower-quality outputs [5]. DMs strike a balance by offering high-quality, diverse images through controlled, stepwise refinement. However, the computational demands of DMs pose a significant limitation, making them less scalable for real-time or large-scale applications [6, 7].

Future research could explore the development of the models that unite the advantages of GANs, VAEs, and DMs to improve both image quality and computational efficiency. In addition, enhancing the interpretability and control of T2I models through user feedback or advanced text encodings could lead to more semantically meaningful outputs. Furthermore, applying T2I models to specialized fields like medical imaging or scientific visualization, where accuracy and detail are critical, presents a promising direction for future work [1].

4. Conclusion

This research introduces a novel method for T2I generation by utilizing detailed textual prompts in combination with a diffusion-based generative framework. The proposed technique skillfully combines CLIP embeddings with DMs to markedly enhance both the visual accuracy and the semantic coherence of the images produced. The process employed in this approach is systematic, starting with the extraction of text embeddings using CLIP, which are subsequently decoded into image embeddings. The final step involves iterative refinement of these images through the diffusion models. Through comprehensive experiments conducted on the MS-COCO dataset, the efficacy of this methodology is clearly demonstrated. Results indicate that integrating CLIP with DMs leads to superior performance compared to more traditional approaches, such as GANs and VAEs, particularly in terms of generating high-quality and diverse images. The diffusion model shows notable advantages, achieving lower FID and higher IS, metrics that reflect its capability to generate realistic as well as semantically consistent images. Moving forward, future investigations will prioritize the creation of hybrid models that merge the strengths of GANs, VAEs, and DMs, aiming to optimize the overall performance of T2I generation systems. Additionally, there will be a concentrated focus on enhancing computational efficiency and delving into specific applications, such as those in the fields of medical imaging and scientific visualization, to further expand the potential use cases and overall impact of T2I generation technology.

References

[1] Baltrušaitis T Ahuja C Morency L P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443

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

[3] Goodfellow I Pouget-Abadie J Mirza M Xu B Warde-Farley D Ozair S Bengio Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27

DOI: 10.54254/2755-2721/105/2024TJ0053

[4] Brock A. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv preprint 1809.11096
[5] Kingma D P. (2013). Auto-encoding variational bayes arXiv preprint 1312.6114
[6] Ho J Jain A Abbeel P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851
[7] Lugmayr A Danelljan M Romero A Yu F Timofte R Van Gool L. (2022) Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11461-11471
[8] Radford A Kim J W Hallacy C Ramesh A Goh G Agarwal S Sutskever I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, 8748-8763
[9] Lin T Y Maire M Belongie S Hays J Perona P Ramanan D Zitnick C L. (2014). Microsoft COCO: Common objects in context. In Computer Vision–ECCV Conference, 740-755
[10] Dhariwal P Nichol A. (2021). Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems, 8780-8794
[11] Zhang E Y Cheok A D Pan Z Cai J Yan Y. (2023). From Turing to Transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models. Science, 5(4), 46