Dean&Francis
ISSN 2959-6157

A Comparative Analysis of Diffusion-Based Models in Text-to-Image Generation

Yiyang Wu

Pinehurst school, Auckland, 0630, New Zealand
Corresponding author: yw30013@pinehurst.school.nz

Abstract:

The recent years have seen the rapid development of artificial intelligence image generation methods with the introduction of a wide range of different models to produce images based on a specific text description. The paper explores the design, functionality, and restrictions of four well-known diffusion-based text-to-image generative artificial intelligence systems: Midjourney, DALL-E2, Stable Diffusion, and Imagen. The paper defines the theoretical context of the diffusion probabilistic model and the manner in which it has been applied in the four models of artificial intelligence. This paper uses the comparative and evaluation performance of these models based on the available empirical research studies on the same in relation to image fidelity, prompt adherence, creativity, and bias on various parameters. The results of the analysis show that Stable Diffusion works well in the generation of photorealistic pictures of the human face, Midjourney works well in creativity and visual image in artistic settings, and Imagen works well in complex text description. These models have also exhibited several challenges and limitations. The findings depict the difficulties in the production of precise, just, and imaginative imagery in a variety of situations.

Keywords: Generative artificial intelligence; Text-to-image generation; Diffusion model; Midjourney; DALL-E 2

1. Introduction

Recent advancements in generative AI have seen the fast creation of text to image models capable of producing realistic images on the basis of textural prompts. Diffusion-based generative models have gained immense popularity as one of the many generative approaches due to their ability to produce images with high accuracy, diversity, and semantic matching. These models are trained to learn to sequentially convert random noise to structured data through the reverse denoising operation, which offers a stable architecture to generate images of high resolution. Midjourney, DALL-E 2, Stable Diffusion and Imagen are based on the diffusion model and have been applied to a variety of applications in art, design, media, and content creation [1,2,3].

Analyzing the architectural design and performance

of these models, Ramesh et al. describes that DALL-E 2 combines Contrastive Language-Image Pre-training (CLIP) with a diffusion prior to matching the textual and visual representations [4]. As described by Rombach et al. [2], the Latent Diffusion Model that is used in Stable Diffusion is much more efficient in terms of calculations as it uses diffusion in a reduced latent space. According to Saharia et al. [3], Imagen applies a large pretrained language model encoder with cascaded diffusion models to promote prompt understanding and image fidelity. Comparative studies also point to differences in models of performance. Borji shows that Stable Diffusion scores high on photorealism in facial generation tasks in terms of Fréchet Inception Distance (FID) scores [5], Ibrahim et al. realize that Midjourney performs better on creativity and visual quality in Emirati architecture design tasks. Regardless of these developments [6], several analyses find that there are still long-standing shortcomings, such as the difficulty in depicting fine details, spatial consistency and the fact that generated images have cultural, racial, and gender biases, which are part of the imbalanced training data and create significant ethical issues [7,8].

Although the image generation has made a leaping advance, current studies tend to emphasize on particular models or unique assessment variables, which leads to a partial knowledge of their respective strengths and weaknesses. The main findings of this paper are the following. Firstly, clarifies theoretical distinctions among diffusion models like DALL-E 2 and Stable Diffusion. Secondly, systematically compares model performance on fidelity, prompt adherence, creativity, and bias. Thirdly, synthesizes findings to outline future research directions in text-to-image generation. The purpose of this paper is to offer a comparative evaluation of the most popular text-to-image generative models, especially diffusion-based ones. First, Section 2 gives the theoretical background of the diffusion models, both forward and reverse processes, and the architecture of representative systems, such as DALL-E 2, Stable Diffusion, and Imagen with major distinctions. Second, Section 3 is a review of recent empirical research to compare model performance in image fidelity, prompt adherence, creativity and bias. Finally, the last part is dedicated to the conclusion where the main findings of the paper have been summarized and future research directions of text-to-image generation have been discussed.

2. Generative AI diffusion-based models

2.1 Diffusion model in Midjourney

Midjourney uses the diffusion probabilistic model, also known as simply diffusion model, in its text-to-image generation [1]. Its typical principle is shown in Fig. 1.

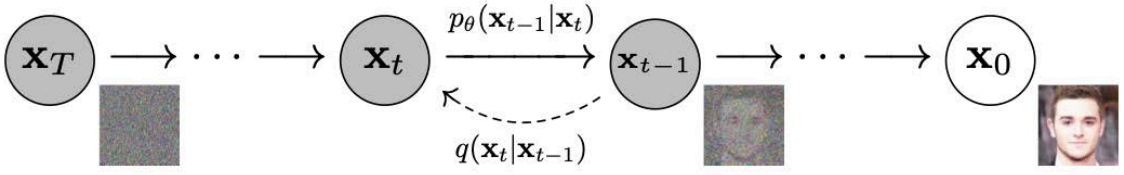

Fig. 1 Illustration of diffusion model [1].

The two major processes involved in the model are a forward diffusion process and a reverse generative process, which are both parameterized Markov chains frameworks. The forward process is a diffusion process and a deterministic algorithm whereby a sequence of small Gaussian noise is sequentially applied to an image, gradually converting the original image to pure Gaussian noise. The forward Markov chain has a predetermined variance schedule, and has no trained parameters, but the reverse process is parameterized and learned. The diffusion process is reversed in the reverse process where an image is formed by the noise. In this case the model is not directly predicting the denoised image but rather predicting the noise which was introduced at each step of the forward process. This is done through training a neural network, most often a U-Net architecture, conditioned on the current timestep. The network is fed an input noisy image and gives the prediction of the noisy component at that level. A comparison between the predicted noise and the true noise that was added is the training objective, frequently in the form of a mean squared error loss. In the sampling process, pure random Gaussian noise is used as the starting point of the generation process. The model will then proceed to apply the reverse transition repeatedly, with its predictions of noise used to gradually predict and eliminate some of the noise at every step. This is how the first noise is gradually converted into a coherent and realistic picture.

2.2 DALL-E 2

DALL E 2 is an Open AI-created model that incorporates

ISSN 2959-6157

the Contrastive Language-Image Pre-training (CLIP) and a diffusion-based generative model, which is a two-stage image generation process [4]. Open AI CLIP is a deep learning model that is trained on about 400 million pairs of images and textual captions, where both the image and the text captions are run through distinct encoders and projected to a common high-dimensional space, and CLIP learns to understand the image and text captions. Cosine similarity is computed between the encodings of the text and image, and the model is trying to maximize the similarity of embeddings of identical image-text pairs, and minimize the similarity of those pairs that do not match. To generate images, an element is used called the diffusion prior transforming CLIP text encoding into CLIP image encoding. This is used to implement this diffusion prior in many cases with a transformer processing the text encoding and producing an image encoding. Then this image encoding is fed into a transformer model created by OpenAI known as GLIDE that produces the final image. GLIDE directly uses cross-attention layers in its UNet architecture, enabling the model to focus on the sequence of text embeddings at every step of the generation process. In this way, the model can guarantee that the produced image is semantically consistent with the original text input during the process of generating it.

2.3 Stable Diffusion

Stable Diffusion uses the Latent Diffusion Model [2], where the diffusion process is run in a reduced latent space instead of pixel space, thus making the computation much more efficient. The core formula in this regard is

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon-\epsilon_\theta\left(z_t,t\right)\right\|_2^2\right] \quad (1)$$

where $\mathbb{E}$ is the expectation, $\mathcal{E}(x)$ denotes the encoder that maps the input x to the latent space, ? denotes the noise sampled from standard normal distribution $\mathcal{N}(0,1)$, $\epsilon_\theta\left(z_t,t\right)$ represents the noise predicted by the neural network with parameters θ at timestep t, and z_t is the noisy latent representation at timestep t.

An autoencoder is first trained to reduce images to a lower dimensional latent code. The encoder part of the system down-samples the input image gradually and the decoder then recovers the original image based on this small latent representation. The autoencoder is trained using perceptual loss to preserve visual quality; an adversarial (patch-based) loss to preserve realism and texture quality; and using either a KL regularization (using the latent distribution to approximate a standard normal distribution) or using vector quantization (VQ). The compression is determined by the down sampling factor f, usually 4 to 8 which is a balance between computational efficiency and the quality of image detail. After this, a diffusion model is trained in the compressed latent space of an autoencoder instead of pixel space. Since the latent representation has significantly smaller spatial dimensions, this method reduces computational cost but still generates similarly as well as pixel-space diffusion models. To use conditional generation, a domain-specific encoder processes the conditioning input, e.g., a text prompt, into an intermediate representation. U-Net architecture of the diffusion model is then expanded by cross-attention layers, which enable the model to build and dynamically optimize relationships between the latent features and the conditioning representation at various steps of the denoising process.

2.4 Imagen

Imagen is a text-to-image model that is proposed by Google DeepMind and is a text encoder based on a large language paired with a cascaded diffusion model [3]. The general architecture consists of three major stages. The initial step is encoding a text using a frozen Large Language Model (LLM). Contrary to most previous methods based on encoders trained on paired image-text data, including CLIP, Imagen is based on a large, pre-trained large language model, the T5-XXL encoder, that stays fixed throughout training. This encoder takes the input prompt and generates a sequence of contextualized embeddings. Imagen uses a large language model to make the image generation process more comprehensive in the sense of better comprehending more intricate prompts. The following step involves the generation of a low-resolution image, usually 64 by 64 pixels, by a conditional diffusion model using the text embeddings. Cross-attention mechanisms built into the U-Net backbone of the model condition the textual input, and the generation process is free to stay tightly coupled with the prompt at any one denoising step. Classifier-free guidance is used to enhance text alignment by enhancing the conditional signal at sampling. Also, dynamic thresholding, a method that produces a mean to stabilize sampling with large guidance weights by adaptively limiting the extreme predicted values, is presented to prevent the generation of over- or undersaturated artifacts. It stabilizes sampling by adaptively limiting severe predicted values, retaining fine detail and photorealistic quality. The last step involves a cascaded super-resolution pipeline, whereby the low-resolution original image is gradually enhanced by two other diffusion models. The former model steps up the resolution to 256×256 pixels and the second model takes another step to get the final resolution to 1024×1024 pixels. Such super-resolution models include

noise conditioning augmentation, whereby the controlled levels of noise are introduced to the low-resolution inputs, to enhance the robustness of the model and allow the models to produce high-quality coherent images with fine detail.

3. Methodology and Models

3.1 Facial Generation

Borji presented a quantitative comparison of three popular text-to-image systems: Stable Diffusion, Midjourney, and DALL-E 2 in Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2 [5]. Borji measured the facial generation of the models by the Fréchet Inception Distance score (FID). Stable Diffusion produces the most realistic faces in general as it has the lowest (best) FID scores in all experiments. In both 5000 faces and 676 faces per model tests, Stable Diffusion is soundly superior to Midjourney and DALL-E 2 in terms of photorealism. Midjourney is more prone to creating surrealistic and anime-aesthetic faces that explains why it has a higher FID score than both Stable Diffusion and DALL-E2. Stable Diffusion had a lower FID score than DALL-E 2. Borji suggests three potential causes of DALL-E 2 performance. First, the DeepFake defenses were deliberately added during training of DALL-E 2 by OpenAI so that it would not memorize those faces that frequently appear on the Internet. Second, instead of faces in crowded, cluttered scenes, DALL-E 2 works better with images that have one point of interest, which makes it better at creating portraits of fictitious individuals than at faces in complex scenes. Third, it could be that the lower performance is due to it having a smaller sample size of only 676 faces to evaluate as FID score is more reliable with a larger sample size. All in all, the FID scores of all three models are greater than those of real images, which is a discrepancy between generated and real faces. Given the low generated quality faces, Borji concludes that the typical problems with all of the models are the inability to produce realistic eyeglasses, challenges with eyeball rendering, failure to produce occluded faces and profile views, and the issue of face symmetry.

3.2 Architectural design generation

The comparison of generated architectural designs of modern villas using the UAE vernacular traditional architecture between Midjourney [6]. DALL-E 2, and Stable Diffusion use image quality, architectural accuracy, prompt adherence. In addition, creativity is described in a comparative and experimental analysis of the leading text-to-image generative artificial intelligence models of residential architectural designs in the region. By Iman Ibrahim, Manar Abu Talib, Just like the results of Borji, Midjourney was the most creative. It also gave the best image quality and a quick compliance, and showed the best overall performance with a final average score of 4.75. DALL-E 2 scored slightly lower than Midjourney, average of 4.13, but it was notably more successful in architectural accuracy even though it was said to be less creative and visually elegant. Stable Diffusion was outperformed by both Midjourney and DALL-E2 by a wide margin, and it had the lowest overall average of 2.75 and uniformly low performance across all evaluation types. Specifically, in immediate compliance and inventiveness its products created are said to be less specific and detailed as those of the other models. On the whole, the research by Ibrahim et al. shows that all three models are limited by the same weakness through their inability to consistently and correctly reflect Emirati architecture and cultural characteristics, and the fact that the most popular model training is not westernized.

3.3 Strengths and Weaknesses Overview

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding Sharia et al (2022) discover that Imagen by Google DeepMind performs better on both automated metrics (FID) and human ratings [3]. Imagen scores 7.27 on the standard COCO benchmark, zero-shot FID-30K, which is considerably better than the score of DALL-E 2 10.39. Likewise, during the conduction of the pairwise human evaluation, raters always favored Imagen to DALL-E 2. In 7 out of 11 categories, Imagen was the preferable option in alignment of images-text and in 11 out of the 11 categories, Imagen was the preferable option in image fidelity. Imagen performs better on spatial relationships among objects, coherent quoted text in images and on complex and lengthy textual descriptions. Nevertheless, Imagen has a number of major weaknesses despite its overall good performance. The most striking is that it is not easy to generate realistic people: the outputs of Imagen were always subject to the preference of human judges when people were not present in the picture. Sharia et al. also discover that Imagen has a general tendency to produce images of individuals with light skin, and the platform is biased towards Western gender stereotypes of some occupations. In addition, in general, the generation of images without humans also shows biases during the representation of activities, events, and objects, which is why the training of common model, as reported by Ibrahim et al., is not influenced by western

ISSN 2959-6157

influences. Also, images such as those of LAION-400M, on which Imagen was trained, have been discovered to be of pornographic nature, racist words, and demotivational words, which raises the question of the capability of the model to always produce harmless content. Some of the weaknesses exhibited by DALL-E 2 include: it is weak in the correct assignment of colours to objects, particularly when there are multiple objects in the prompt; it cannot produce coherent text in images; it is not good at describing spatial relationships correctly and prompts with multiple objects; it is not good at longer and more complex descriptions.

3.4 Bias in Models

Imagining the Far East: Exploring Perceived Biases in AI-Generated Images of East Asian Women Lan and coworkers performed a user-based audit of Midjourney, DALL-E and Stable Diffusion, in which users were tasked with creating images of an attractive East Asian woman with each model, and to rate their impressions of bias [8]. The research established that image satisfaction was only observed to be satisfactory in 43.7% of the generated images. Midjourney recorded the highest satisfactory images rate of 53.7, DALL-E was at 49.6, and Stable Diffusion at 14.4. Overall, the participants were the most satisfied with Midjourney with an average satisfaction of 72.2 out of 100. The mean satisfaction score of 62.8 out of 100 was given to DALL-E and the lowest mean score of 52.3 out of 100 was given to Stable Diffusion. Although Stable Diffusion had fewer dominating biases than DALL-E, it tended to represent East Asian women in a very revealing manner, dull-eyed, and thin by default [9,10]. In their study, Lan and colleagues came up with 18 perceived biases divided into four patterns namely Westernization, overuse or misuse of cultural symbols, sexualization and feminization, and racial stereotypes. Among them, DALL-E was most commonly attributed with Westernization biases, Midjourney with biases related to the overuse or misuse of cultural symbols, and Stable Diffusion with biases related to sexualization and feminization and racial stereotyping [11].

1. Conclusion

A comparative analysis of text-to-image diffusion-based models of text-to-image generation has been provided in this paper. The paper has demonstrated the strengths and weaknesses of the various models through the analysis of theoretical basis of the diffusion-based models and the variations in the designs of the different models such as Midjourney, DALL-E2, Stable Diffusion, and Imagen. The discussion above demonstrates that diffusion-based models are strong and can be applied to create high-quality images, although they vary in their implementation, which results in different strengths and weaknesses of each model. Based on the empirical comparisons in this paper, Stable Diffusion is very good at photorealism in face generation, Midjourney is well-performing in terms of creativity and stylistic variety, Imagen demonstrates good performance with complex prompts and preserving semantic consistency, and DALL-E2 demonstrates balanced performance but does not comprehend complex scenes well and rendering fine details. These distinctions are manifestations of the effects of architectural designs in terms of latent diffusion, large language model encoders, and conditioning.

The analysis also finds that there are some consistent issues with the images being generated by the models. Works of all systems generated are less realistic than real images. The presence of cultural, racial, and gender biases is also a typical restriction that can be observed, which shows the impact of training data and poses serious ethical issues related to fairness and representation. These results indicate that both model design and dataset diversification should be enhanced to guarantee more credible and fair results.

Further studies are needed on how to make complex compositional tasks more robust, and more robust assessment frameworks that combine technical performance with societal impact. The issue of the lack of diversity in training data should be addressed because the existing models tend to be Western-centric. Future research must then focus on the integration of more global datasets in order to minimize the cultural and demographic bias in the produced work.

References

[1] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. Virtual, December 6-12, 2020, 2020.

[2] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2022, 10684-10695.

[3] Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, Ghasemipour SKS, Gontijo Lopes R, Karagol Ayan B, Salimans T, Ho J, Fleet DJ, Norouzi M. Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems. New Orleans, LA; November 28-December 9, 2022, 2022:36479-36494.

[4] Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with CLIP latents.

Preprint. Posted online April 12, 2022. arXiv:2204.06125.
[5] Borji A. Generated faces in the wild: quantitative comparison of Stable Diffusion, Midjourney and DALL·E 2. Preprint. Posted online October 2, 2022. arXiv:2210.00586.
[6] Ibrahim I, Abu Talib M, Ammar A, Tabet Aoul KA, Abuimara T. Comparative and experimental analysis of leading text-to-image generative artificial intelligence models for regional residential architectural designs. Results Eng. 2026, 29:108835.
[7] Lan X, An J, Guo Y, et al. Imagining the Far East: exploring perceived biases in AI-generated images of East Asian women. Preprint. Posted online April 8, 2025. arXiv:2504.04865.
[8] Xu M Y. Application of AI Drawing Technology in the Field of Digital Illustration and Its Process. Tomorrow's Fashion, 2025, (04): 164-166.
[9] Liu F L, Yang L Y. AI Drawing Empowers the Development of 3D Design: Taking Stable Diffusion as an Example. Kunming Metallurgy College Journal, 2025, 41(01): 61-69.
[10] Wang L T. Application of AI Drawing in Textile Pattern Design. Shanghai Apparel, 2024, (12): 19-21.
[11] He J L. Generative AI Drawing Technology Empowers Journalism: Opportunities, Risks, and Mitigation. News Editing, 2024, (05): 125-126.