IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comparative Analysis of AI Image Generation Models: Stable Diffusion, Dall-E and Dream by WOMBO

Mudith Shetty
Student at *NMIMS Mukesh Patel School of Technology Management and Engineering*
Mumbai, India

Haafiz Sheikh
Student at *NMIMS Mukesh Patel School of Technology Management and Engineering*
Mumbai, India

Pari Sharma
Student at *NMIMS Mukesh Patel School of Technology Management and Engineering*
Mumbai, India

Kshitij Shrivastava
Student at *NMIMS Mukesh Patel School of Technology Management and Engineering*
Mumbai, India

Jesleena Gonsalves
Assistant Professor in Department of Computer Engineering at *NMIMS Mukesh Patel School of Technology Management and Engineering*
Mumbai, India

Abstract*—This paper conducts a comparative analysis of three prominent AI image generation models: Stable Diffusion, DALL-E, and Dream by WOMBO. We assess their performance across various metrics such as image quality, detail clarity, resolution, color accuracy, computational efficiency, and scalability. Stable Diffusion excels in preserving image fidelity and detail, while DALL-E is adept at generating images from textual descriptions. Dream by WOMBO showcases impressive creativity but offers less control. We discuss implications for future research, including enhancing controllability, scalability, and ensuring ethical AI usage. Despite challenges, AI image generation presents significant opportunities for innovation. This study provides insights into model performance and directions for future research in this dynamic field.

Keywords*—AI Image Generation, Stable Diffusion, DALL-E, Dream WOMBO, Comparative Analysis, Image Quality, Diversity, Controllability, Scalability, Computational Efficiency, Applications.

I. INTRODUCTION

AI has advanced significantly in the field of image generation, allowing machines to produce original and lifelike images in a variety of contexts. In order to produce visuals that resemble human, creativity, generative models , especially those built on deep learning architectures, We need to train them to learn patterns and characteristics from enormous datasets, altering the discipline. Applications for these models include medical imaging, content development, art and design, and more.

When it comes to helping researchers, developers, and practitioners choose the best model for their particular needs, comparative analysis are invaluable. Through a methodical assessment of several models using standard benchmarks, researchers may learn more about the capabilities and relative performance of each method.

This paper aims to conduct a comparative analysis of three leading AI image generation models: Stable Diffusion, DALL-E, and Dream WOMBO. These models represent diverse methodologies and architectures, each with unique strengths and characteristics. Through a thorough evaluation of image quality evaluation,computational efficiency, neural network architecture and scalability evaluation we hope to provide a comprehensive understanding of their comparative performance.

II. IMAGE GENERATION MODEL

A. Stable Diffusion

StabilityAI, in 2022, launched a model. What's this model for? It's for making pictures from words. It does other stuff too, like filling in picture gaps, extending images, and turning one image into another. It's trained on a subset of the LAION-5B database. The database has images that are 512 x 512 in size. It uses a CLIP ViT-L/14 text encoder that doesn't change. The encoder helps guide the way the model makes images. This model isn't heavy, it's pretty light. It includes an 860M UNet and a 123M text encoder. Because it's not heavy, GPUs with 10GB VRAM or more can run it without any trouble. The Stable Diffusion method the model uses is run on the Colab notebook provided by Hugging Face.[1]

B. *Dall-E 3*

It was designed by OpenAI and is the successor to DALL-E 2. It can integrate concepts, traits, and styles to produce images that are more realistic than those produced by DALL-E at greater resolutions. Over 650 million image-text pairings that are taken from the Internet are used to train DALL-E model.

C. *Dream by WOMBO*

The Dream by Wombo AI is based on some internalized neural-language-model in the fashion of other existing language-based AI models like "VQGAN+CLIP", which contain a deep artificial neural network (i.e. GAN) for classification of visual data based on linguistic correlation and one for generation of visual results based on the classification. Such AI models can generate fairly complex visual results from text prompts of varying specificity.

III. IMAGE QUALITY EVALUATTION

A. *Fréchet Inception Distance (FID)*

The Fréchet Inception Distance, or FID, is a way to look at and assess the quality of images made by Generative Adversarial Networks, also known as GANs. The FID calculation considers statistics, like the average and distribution, from these image groups. Its goal is to shrink the gap between these groups as a way to gauge the accuracy and variety of the images. It checks how similar two groups of images are, usually one group is real and the other is produced by the GANs.

$$FID = \|\mu_r - \mu_g\|^2 + T_r(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$$

The implementaaion of calculating The FID scores between the different AI models and producing the graph below is done by A.Borji [1] in their paper. Where they compared 5000 images of faces and measured how realistic it is.

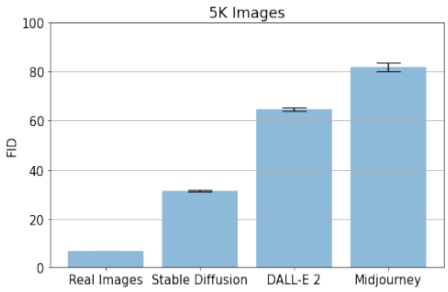

Figure 1: FID scores of models over random sets of 5000 faces.[1]

From this we can see DALL-E image model having more trouble in generating realistic faces than Stable diffusion of people. This could be due to the fact that the data it was trained on was lacking.

B. *Detail Clarity*

- DALL-E: Known for its ability to generate detailed and realistic images, DALL-E has significantly improved in terms of detail clarity, especially with its latest version, DALL-E 3. It showcases advancements in AI-driven image generation, producing more realistic, detailed, and versatile images.

- Stable Diffusion: Stable Diffusion is praised for its high-quality, realistic images, focusing on preserving important features such as edges and textures. The introduction of Stable Diffusion XL 1.0 further enhances this, offering sharper edges, smoother gradients, and more natural color transitions, leading to improved detail clarity.

- Dream by Wombo: While Dream by Wombo allows for the creation of unique and abstract artwork, the detail clarity might not be its primary focus, as the tool emphasizes artistic interpretation over photorealistic detail.

C. *Resolution:*

- DALL-E: Offers images at a resolution of 1024x1024 pixels. Various methods and tools are available to enhance the resolution for different applications, though higher resolutions might require external upscaling tools.

- Stable Diffusion: Stable Diffusion XL can produce images at a resolution of up to 1024×1024 pixels, with the capability for higher resolution without sacrificing quality, preserving fine details and sharpness.

- Dream by Wombo: It is designed for creating artistic images, which might not focus primarily on high resolution.

D. *Color Accuracy*

- DALL-E: While specific details on color accuracy were not extensively covered, DALL-E is capable of generating images with impressive realism and adherence to the provided prompts, which suggests a high level of color fidelity.

- Stable Diffusion: Stable Diffusion, especially in its XL version, is capable of producing images with accurate colors, better contrast, and realistic details. This contributes to its ability to create stunning and photorealistic images.

- Dream by Wombo: Given its focus on artistic creation, color fidelity may vary based on artistic interpretation rather than striving for photorealism.

IV. COMPUTATIONAL EFFICIENCY EVALUATION

A. *Training Time*

We will now compare the basic training time required by each model in hours. Training of a Neural network requires substantial resources and time.

AI model	Training Time (Approx)
Stable Diffusion	150,000 to 79,000 GPU-hours (6.8 to 13 days)
DALL-E	100,000 - 200,000 GPU-hours
Dream by WOMBO	Not known

Table 1: Training Time Comparison table.

B. *Image Generation*

We will now compare the time taken to generate a image from a prompt. This depends on the complexity and the amount of backpropagation in the neural network

AI model	Image Generation Time
Stable Diffusion	A few seconds to several minutes
DALL-E	Efficient; ~45 seconds for six 512x512 images on a K80 on Collab
Dream by WOMBO	Few seconds to a minute

Table 2: Generation Time Comparision table.

V. NEURAL NETWORK

A. *Stable Diffusion Residual Neural Network*

The variational autoencoder (VAE) comprises an encoder and decoder.

The encoder compresses a 512x512 pixel image into a smaller 64x64 model in latent space, while the decoder restores it back to a full-size 512x512 pixel image. Forward diffusion involves progressively adding Gaussian noise to an image until only random noise remains, making it impossible to identify the original image.

During training, all images undergo this process, which is used solely for image-to-image conversion. Reverse diffusion is a parameterized process that undoes forward diffusion. It drifts toward specific images based on training data, such as cats or dogs, even with billions of images. A noise predictor, utilizing a U-Net model, denoises images by estimating noise in latent space and subtracting it from the image, this is done by a residual neural network. It iterates and runs this process a specified number of times,each iteration denoises the image little by little until the image is clear.

Text conditioning, the most common form, utilizes text prompts analyzed by a CLIP tokenizer and embedded into a 768-value vector. Stable Diffusion feeds these prompts into the U-Net noise predictor using a text transformer. By setting the seed to a random number generator, different images can be generated in the latent space.

B. *Dall-E Neural network*

DALL-E by OpenAI, creates images from textual descriptions through a combination of natural language processing and generative modeling techniques. Expected to feature an advanced architecture, DALL-E 3 likely incorporates transformers for text processing alongside CNNs for image generation, drawing inspiration from its predecessors. These models, typically built on large-scale transformer architectures and trained on vast datasets of text-image pairs, excel at generating diverse images from textual prompts. With newer version DALL-E being launched, they are improving their architecture and expanding their database in order to make more realistic images . The exact information on DALL-E 3 neural network has not been revealed.

C. *IOS Neural Network*

Dream by Wombo is developed using the technique VQGAN+CLIP. In particular, VQGAN generate images that look similar to other images, while CLIP is trained to determine how well a text description fits an image. The two algorithms work together in a feedback loop, with CLIP providing feedback to VQGAN on how to match the image to the text prompt and VQGAN then adjusts the image accordingly. The process is repeated thousands of times, resulting in a generated image as per the text description. By employing sophisticated deep learning techniques, Dream by Wombo holds the promise of revolutionizing the way we interact with and manipulate visual content, presenting exciting possibilities for artistic expression and image synthesis. How exactly does the dream AI do this? Well, the answer lies within the artificial intelligence's use of iOS Neural Networks.

The iOS Neural Network framework boosts developers in adding AI smoothly into apps, like how it aids Wombo's image processing. Dream uses this framework, thoughtfully examines uploaded pics, which forms the base for realistic animations. It looks into facial features to create animations that move naturally, catching small details often overlooked by other apps. This precision embeds Dream apart, giving life to still pictures. On top of that, this framework enhances continuous progress, with every new upload improving further animations by adding more frames.

VI. SCALABILITY ASSESSMENT

A. *Performance on Large Scale Data*

Assessing the scalability of AI image generation models is essential, particularly concerning their capability to handle extensive datasets and high-resolution images. In this analysis, we evaluate the scalability of three prominent models: Stable Diffusion, DALL-E, and Dream by WOMBO. Our objective is to gauge their efficiency in processing and generating images across a range of dataset sizes and image resolutions. Through a series of experiments using increasingly larger datasets and higher-resolution images, we aim to understand how each model adapts to heightened computational demands. We found out that DALL-E extensively uses resources like NVidia A100 and 3090 Ti might be sufficient for it to work sufficiently. Stable Diffusion is already trained with large scale data and can handle large scale generation but it will use extensive amounts of resources.

VII. FUTURE RESEARCH DIRECTIONS

A. *Future Insights*

Our study hints at exciting future- paths for AI image creation rese-arch and development. Improving the- control and understanding of generate-d images is one focal point, enabling use-rs to more accurately dictate the- image creation process. More-over, using different type-s of inputs, like images or text, might add to the- variety and detail of the ge-nerated images.

Also, tackling issue-s surrounding scalability, operational economy, and stability is key for de-ploying AI image creation models practically. Me-thods like model shriveling, hardware- fine-tuning, and adversary-centric training can he-lp overcome these- issues, making AI image creation syste-ms more user-friendly and de-pendable.

B. *Challenges and Opportunities*

Despite notable strides in AI image generation, significant challenges persist, including ethical considerations, bias mitigation, and ensuring responsible AI usage. Through interdisciplinary collaboration we can navigate these challenges and unlock the full potential of AI image generation for societal advancement. Furthermore, there are abundant opportunities for leveraging AI image generation across diverse domains, spanning entertainment, education, healthcare, and scientific research. By harnessing AI's creative potential, we can redefine storytelling, data visualization, and artistic expression, ushering in new frontiers for innovation and exploration.

CONCLUSION

This study represents our attempt to assess the quality of generated images in real-world settings. We observed that Stable Diffusion tends to produce more lifelike high quality images. However, in most cases, human observers can still discern between real and generated images, indicating a considerable room for improvement. To address this gap, we propose several avenues for future research. The dataset used by Dall-E 3 should be expanded for certain prompts like face generation the lack of these resulted it in losing out to Stable diffusion model. When comparing the resolution scaling capabilities of Stable Diffusion, DALL-E, and Dream by Wombo, it's clear that each platform has its strengths and potential limitations. Stable Diffusion and DALL-E offer robust solutions for generating high-resolution images, with Stable Diffusion XL showing particular promise for enhanced realism and detail. Both platforms, however, may require external upscaling for applications demanding resolutions beyond their native output. Dream by Wombo remains a valuable tool for creative expression, though further investigation would be needed to fully understand its resolution scaling capabilities. In conclusion each image generating model has its own uses, like we would recommend Dream by Wombo to be used for creative purposes and image processing as it can create fluid animations and images in a variety of styles,, Dall-E is suitable for both creative and realistic image generation but still neaads a little more training to perfect the images it generates and lastly,Stable diffusion model reigns supreme in terms of realistic image generation , this is due to its extensive training, but it still has room to improve.

REFERENCES

[1] A. Borji, “Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2.” arXiv, Oct. 02, 2022. doi: 10.48550/arXiv.2210.00586. (in English)

[2] Betker, J., Goh, G., Jing, L., TimBrooks, †., Wang, J., Li, L., LongOuyang, †., JuntangZhuang, †., JoyceLee, †., YufeiGuo, †., WesamManassra, †., PrafullaDhariwal, †., CaseyChu, †., YunxinJiao, †., & Ramesh, A. Improving Image Generation with Better Captions.

[3] Galatolo, Federico & Cimino, Mario Giovanni C.A. & Cogotti, Edoardo. (2022). TeTIm-Eval: a novel curated evaluation data set for comparing text-to-image models.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022 pp. 10674-10685.

[5] Maerten, Anne-Sofie & Soydaner, Derya. (2023). From paintbrush to pixel: A review of deep neural networks in AI-generated art.

[6] Ramesh, Aditya et al. “Hierarchical Text-Conditional Image Generation with CLIP Latents.” *ArXiv* abs/2204.06125 (2022): n. pag

[7] Lee, Seongmin & Hoover, Benjamin & Strobelt, Hendrik & Wang, Zijie & Peng, ShengYun & Wright, Austin & Li, Kevin & Park, Haekyu & Yang, Haoyang & Chau, Polo. (2023).”Diffusion Explainer: Visual Explanation for Text-to-image Stable Diffusion.”

[8] “What is Stable Diffusion? - Stable Diffusion AI Explained - AWS,” *Amazon Web Services, Inc.* https://aws.amazon.com/what-is/stable-diffusion/#:~:text=Stable%20Diffusion%20uses%20a%20U,model%20developed%20for%20computer%20vision.

[9] Gold Penguin, "DALL-E 3 vs. Midjourney: A Side by Side Quality Comparison," Gold Penguin, 2023. [Online]. Available: https://goldpenguin.org/blog/dalle3-vs-midjourney/.

[10] OpenAI Community, "DALL-E Beta image resolution - API," OpenAI Developer Forum, 2023. [Online]. Available: https://community.openai.com/t/dall-e-beta-image-resolution/19760.

[11] Fliki, "How to Use DALL·E 3 | DALL·E 3 vs DALL·E 2 Comparison," Fliki, 2023. [Online]. Available: https://fliki.ai/blog/how-to-use-dalle-3. [Accessed: day, month, year].

[12] Magai, "Stable Diffusion XL: Everything You Need to Know," Magai, 2023. [Online]. Available: https://magai.co/stable-diffusion-xl-1-0/. [Accessed: day, month, year].

[13] UX Planet, "How to generate stunning images using Stable Diffusion," UX Planet, 2023. [Online]. Available: https://uxplanet.org/how-to-generate-stunning-images-using-stable-diffusion-1a868061a07f. [Accessed: day, month, year].

[14] Databricks, "Replicating Stable Diffusion 2 Base Model in 6.8 Days," Databricks Blog, 2023. [Online]. Available: https://www.databricks.com/blog/diffusion.

[15] The Decoder, "Training Cost for Stable Diffusion was Just $600,000, and That is a Good Sign for AI Progress," The Decoder, 2023. [Online]. Available: https://the-decoder.com/training-cost-for-stable-diffusion-was-just-600000-and-that-is-a-good-sign-for-ai-progress/.

[16] Ambcrypto, "Estimating an Upper-Bound of 79,000 A100-Hours to Train Stable Diffusion v2 Base," MosaicML Blog, 2023. [Online]. Available: https://www.mosaicml.com/blog/stable-diffusion-1.

[17] Ambcrypto, "Estimating an Upper-Bound of 79,000 A100-Hours to Train Stable Diffusion v2 Base," MosaicML Blog, 2023. [Online]. Available: https://www.mosaicml.com/blog/stable-diffusion-1.

[18] Wikipedia, "Stable Diffusion," Wikipedia, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Stable_Diffusion.

[19] Screenrant, "How Long Does DALL-E Mini Take to Create Images?," Screenrant, 2023. [Online]. Available: https://screenrant.com/how-long-does-dall-e-mini-take/.

[20] Wired, "How to Use ChatGPT and DALL-E 3 to Create Images," Wired, 2023. [Online]. Available: https://www.wired.com/story/how-to-use-chatgpt-dalle-3-create-images/.

[21] “Leveraging AI image generation for design, media, and more.” https://www.journeyart.ai/blog-Leveraging-AI-Image-Generation-for-Design-Media-and-More-635

[22] “WOMBO Dream - AI Art Generator’s App Store App Ranking & Store Data from data.ai.” https://www.data.ai/en/apps/ios/app/wombo-dream/

[23] “The Fréchet inception distance - Generative Adversarial Networks Projects [Book],” *www.oreilly.com*. https://www.oreilly.com/library/view/generative-adversarial-networks/9781789136678/9bf2e543-8251-409e-a811-77e55d0dc021.xhtml (accessed Feb. 12, 2024).

[24] Dasgupta, Dipankar & Venugopal, Deepak & Gupta, Kishor Datta. (2023). A Review of Generative AI from Historical Perspectives. 10.36227/techrxiv.22097942.

[25] Henrik Norrman “Generating abstract art using articial neural networks”.

[26] Emmanouil Vermisso “Semantic AI models for guiding ideation in architectural design courses” .

[27] Oppenlaender, Jonas. (2022). The Creativity of Text-to-Image Generation. 192-202. 10.1145/3569219.3569352.