Custom Concept Text-to-Image Using Stable Diffusion Model in Generative Artificial Intelligence

Alam Rahmatulloh[1]*

[1]Department of Informatics, Siliwangi University, Tasikmalaya City 46115 (Indonesia)

* Corresponding author: alam@unsil.ac.id

Received 10 February 2025 | Accepted 20 April 2025 | Early Access 3 Mei 2025

Abstract

The ability of algorithms to produce content that closely mimics human work has revolutionized several fields thanks to generative artificial intelligence, or Gen AI. However, these developments also raise questions about generative models' transparency, predictability, and behavior. Considering the relevance of this topic and the expanding influence of AI on society, research into it is imperative. This paper aims to empirically explore the nuances of behavior in the setting of discriminative generative AI, using the stable diffusion model as an example. We will be better equipped to handle obstacles and guarantee the ethical and responsible application of generative AI in a world that is changing quickly if we have a deeper grasp of this phenomenon. The research method is carried out in several stages, such as dataset collection, modeling, testing, and analysis of results. The research results show that generative artificial intelligence can create realistic images like the original. However, there are still several challenges, including the availability of a reasonably large dataset for training data and high and long computing times. Likewise, the results of the Fréchet Inception Distance (FID) test were still quite large, namely 1284.4430, which shows that the quality of this model is still not good.

Keywords

Custom Concept, Generative Artificial Intelligence, Stable Diffusion, Text-to-image

I. Introduction

Technology is still evolving, and people's daily lives are incorporating more artificial intelligence. The influence of AI is pretty real, ranging from recommendation algorithms to virtual assistants. Simultaneously, generative AI models' predictability and comprehension of their behavior present challenges. Given the unpredictability of generative AI and the possible social effects of the content it generates, the findings of this study should be closely examined to verify that they adhere to ethical standards, legal requirements, and societal norms. Furthermore, because generative models conflate synthetic and real data, their discriminativeness and selectivity also raise particular issues. Complex patterns and relationships in data can be recognized and understood by neural networks [1]. Neural networks, unlike traditional rule-based systems, have the ability to automatically recognize these patterns from the training data without requiring explicit programming or human understanding of the underlying mechanics. The trait of neural networks that is sometimes referred to as the "black box" property is being described here [2].

Analysts can also process large data sets, but they cannot identify hidden patterns or create multidimensional relationships due to cognitive constraints. Conversely, neural networks thrive at identifying minute details that may be hard to notice or beyond human comprehension and recording intricate relationships. They can recognize complicated correlations between input data features, identify nonlinear relationships, and recognize hierarchical structures.

Neural networks possess the advantage of obtaining abstract data representations. Due to their expertise in managing data with a large number of dimensions, they are able to evaluate complex information and identify nuanced patterns. Each layer inside a neural network acquires increasingly complex representations. Neural networks consist of interconnected nodes organized in layers, which are capable of learning to transform incoming data into hierarchical arrays of features [3]. Thanks to their hierarchical representation, they can recognize and comprehend patterns that might not be immediately apparent to humans.

Neural networks are opaque systems that pose many challenges. Due to its complex and non-linear structure, it is difficult to understand the rationale behind the network's

Please cite this article as: Rahmatulloh, A. (2025). Custom Concept Text-to-Image Using Stable Diffusion Model in Generative Artificial Intelligence. International Journal of Informatics and Computing, 1(1), 1-10.

specific judgments or predictions. Explainability is crucial in sensitive fields like banking and healthcare, where the absence of interpretability raises worries [4]. Therefore, empirical investigations of neural networks, including generative networks, can help better understand the nature of their work [5].

There are a thousand words in a picture. As they say, pictures convey a message more effectively than words alone. When people read a story in text, their imaginations can conjure pertinent pictures, enhancing comprehension and enjoyment. As a result, creating an automated system that uses texture descriptions to create visually realistic images—a process known as text-to-image tasks—is challenging and represents a significant advancement towards artificial intelligence that is more like human intelligence [6]–[9]. Text-to-image tasks have become one of the most spectacular applications in computer vision [11]–[23] with the emergence of deep learning [10].

The term "artificial intelligence" now encompasses a wide range of models. They play a vital role in developing many talents such as robotics, computer vision, machine learning, natural language processing, comprehension, and generation. Artificial intelligence can be categorized into two main types: narrow AI and general AI [24]. Systems that are intelligent enough to understand, learn, and apply knowledge from a wide range of fields are called general artificial intelligence (AI). At this point, general artificial intelligence is still unrealized and only a theory.

Conversely, narrow AI is made to accomplish particular jobs within limited capabilities. Voice assistants, image recognition software, recommendation engines, etc., are a few examples of narrow artificial intelligence. Let's discuss generative artificial intelligence in greater detail.

II. Literature Review

A. Generative AI and its Variations

A branch of artificial intelligence called "generative AI" aims to create data or content that seems to be made by humans. This system generates fresh data that exhibits patterns comparable to the training data by utilizing machine learning capabilities, particularly generative models [25]. Many kinds of models are being created and applied concurrently in generative AI.

1. *Generative Adversarial Networks (GAN)*

   A generator and a discriminator are the two essential parts of a GAN that are trained simultaneously. While the discriminator separates authentic samples from falsely generated samples, the generator attempts to produce realistic data samples with text or images. These two elements engage in competitive interaction, and the generator eventually can produce ever-more-believable content as it learns;

2. *Model transformation*

   The GPT model is one illustration. They use self-attention mechanisms to identify subtle and complex correlations present in the incoming material. These models have been applied with success to many different generative applications, including text generation, image synthesis, and music composition;

3. *Variational autoencoders (VAEs)*

   This generative model has the ability to process input by both encoding and decoding it. A Variational Autoencoder (VAE) normally comprises two main components: a decoder, which accurately reconstructs the input data from the latent space, and an encoder, which transforms the input data into a representation of the latent space. Through the extraction and decoding of points from the latent space, VAE facilitates the generation of novel samples;

4. *Model autoregressive*

   Probabilistic models, known as autoregressive models, sequentially produce new data, with each element conditionally reliant on the previous one. Examples of autoregressive models that can produce intricate and contextually relevant textual material are the GPT-3 and GPT-4 language models.

5. Deep Reinforcement Learning

   Deep reinforcement learning combines deep learning techniques with reinforcement learning algorithms, enabling the creation of artificial intelligence. These models excel at engaging with the environment and receive either rewards or punishments according to their actions. By taking calculated risks and strategically optimizing behaviors, individuals can develop novel patterns of behavior or tactics that result in higher rewards.

   The models discussed here constitute a small portion of the wide array of generative AI models. Scientists and engineers are currently investigating various methods to create models capable of generating a wider variety of unique material in multiple fields.

B. Historical Development of Generative AI Models

These advances pave the way for a bright future in the field of generative AI, promising further innovation and breakthroughs.

1. *In 2012*

   The inception of AlexNet occurred in 2012. The authors of this study, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, introduced this model called "ImageNet Classification with Deep Convolutional Neural Networks." AlexNet has contributed to the renewed interest in deep learning and convolutional neural networks (CNNs), particularly following its remarkable victory in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition.

2. *In 2013*

   The introduction of Variational Autoencoders (VAEs) occurred in 2013. This model was introduced by Diederik P. Kingma and Max Welling in a paper entitled "Auto-Encoding Variational Bayes". VAEs are a kind of generative models that employ variational Bayesian learning to integrate probabilistic modeling and neural networks, allowing for the creation of novel and authentic data from latent distributions.

3. In 2014

The concept of Generative Adversarial Networks (GANs) was initially proposed by Ian Goodfellow and his colleagues in 2014. GANs are a specific kind of artificial neural network models that comprise of two primary components: a generator and a discriminator. These components engage in a competitive learning process. The generator attempts to generate new data that closely resembles the original training data, while the discriminator aims to distinguish between the original data and the data produced by the generator.

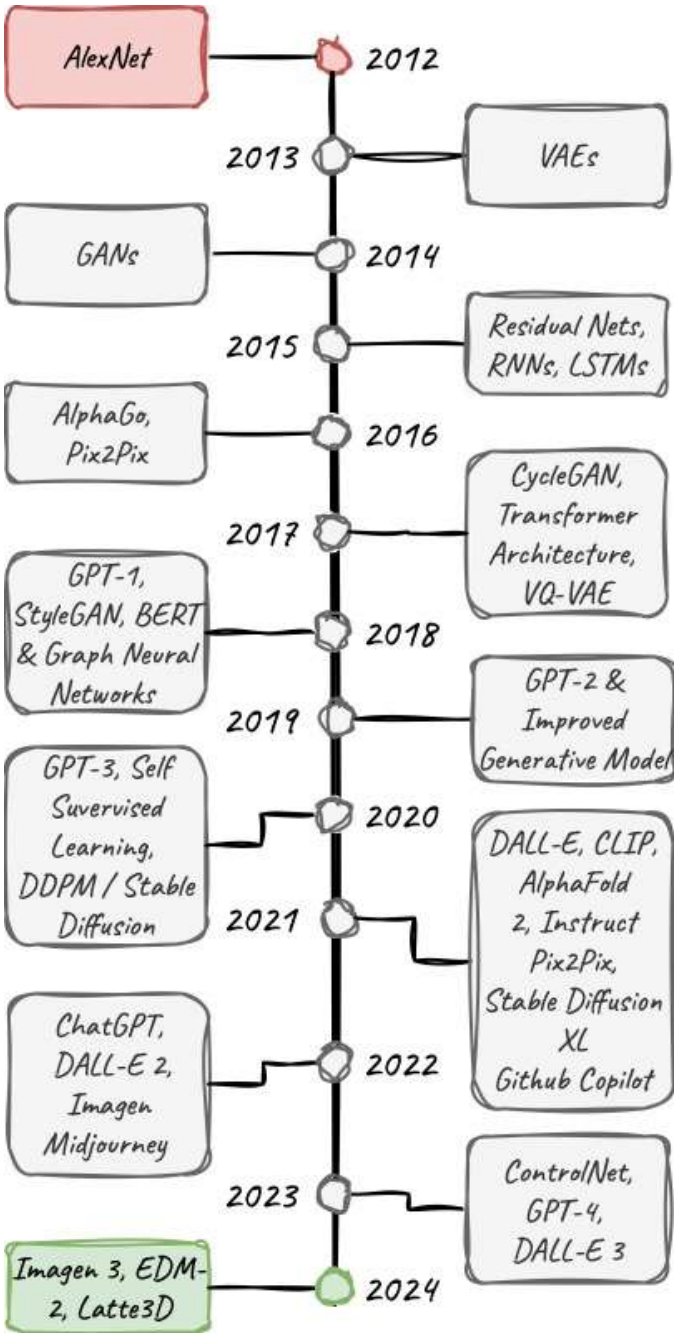

Fig. 1. Evolution of Generative AI Models

4. In 2015

Residual Networks, also known as ResNets, were first presented in 2015 by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun from Microsoft Research. Their renowned paper titled "Deep Residual Learning for Image Recognition" was published at the European Conference on Computer Vision (ECCV) in 2016, despite the fact that the study had been conducted before. Residual Networks (ResNets) were formally introduced in around 2015.

5. In 2016

AlphaGo, an AI program created by DeepMind (a subsidiary of Alphabet Inc.), was initially released to the public in 2016. In March 2016, AlphaGo emerged victorious over professional Go player Lee Sedol in a series of widely observed matches. Global. This triumph signifies a notable progression in the realm of artificial intelligence, given that Go is often regarded as one of the most intricate games in existence. This accomplishment demonstrates the capacity of computers to tackle increasingly intricate and conceptual challenges.

The Pix2pix method, developed by Phillip Isola and his colleagues in 2016, utilizes Generative Adversarial Networks (GANs) to create a mapping between two images. This research makes substantial contributions to the field of image processing and machine learning, particularly in the area of image transformation, such as converting an image from one domain to another (e.g., converting a sketch picture to a color image).

6. In 2017

In 2017, Jun-Yan Zhu and his colleagues unveiled CycleGAN, an advancement of Generative Adversarial Networks (GANs) that enables unsupervised learning for mapping pictures between different domains. This research expands upon the concept of Generative Adversarial Networks (GANs) by incorporating the concept of cyclic mapping. This enables the model to transfer information between different domains without the need for explicitly labeled training data sets.

The Transformer design was first introduced in a paper titled "Attention is All You Need" by Vaswani et al., which was presented at the Neural Information Processing Systems (NeurIPS) conference in 2017. The Transformer architecture is a significant breakthrough in the domains of natural language processing and image identification. It has served as the foundation for numerous advancements in language models and image recognition models, including BERT, GPT (Generative Pre-trained Transformer), and other notable examples. The Transformer architecture was formally announced in 2017.

The VQ-VAE (Vector Quantized Variational Autoencoder) was initially presented in a research paper titled "Neural Discrete Representation Learning" authored by van den Oord et al. and published in 2017. This study presents a novel method for discrete representation learning in variational autoencoders (VAE). It utilizes a vector quantization training code to acquire a discrete representation of the data. VQ-VAE was initially introduced in 2017.

7. In 2018

OpenAI introduced GPT-1 (Generative Pre-trained Transformer 1) in June 2018. The initial model in the GPT series employs the Transformer architecture and has

undergone pre-training using an extensive dataset to execute text-related tasks such as automated text synthesis, translation, and other forms of natural language comprehension. Consequently, GPT-1 was introduced in 2018.

Tero Karras and colleagues from NVIDIA presented StyleGAN in December 2018. StyleGAN is a method for generating realistic images using Generative Adversarial Networks (GANs). The primary objective of StyleGAN is to generate high-quality images with enhanced controllability, allowing for precise manipulation of specific styles or qualities in the generated images. StyleGAN was officially unveiled in December 2018.

The BERT model, also known as Bidirectional Encoder Representations from Transformers, was first presented in 2018 by Jacob Devlin and his colleagues from Google AI Language. The primary publication that introduced BERT, titled "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," was released in May 2019. Prior to the publication of this work, the Google research team had already internally developed and introduced the BERT model and related concepts. Officially, BERT was launched or introduced in 2018.

8. In 2019

The introduction of GPT-2 (Generative Pre-trained Transformer 2) by OpenAI took place in February 2019. This model is an iteration of GPT-1 that incorporates a larger Transformer architecture and enhanced capabilities for generating lengthier and more authentic text. GPT-2 underwent training on a more extensive dataset compared to GPT-1, enabling it to achieve superior proficiency in tasks related to natural language processing. GPT-2 was released in February 2019.

9. In 2020

OpenAI introduced GPT-3 (Generative Pre-trained Transformer 3) in June 2020. The GPT models in this series utilize an advanced Transformer architecture with an extensive number of characteristics, making it the most recent version. GPT-3 has exhibited remarkable proficiency in producing text that closely resembles human-level language across a range of natural language processing tasks. GPT-3 was introduced in June 2020.

The article named "Improved Denoising Diffusion Probabilistic Models" by Jonathan Ho et al. introduced DDPM (Diffusion Probabilistic Models) / Stable Diffusion. It was published at the International Conference on Machine Learning (ICML) in 2020. Officially, the DDPM was implemented in 2020.

10. In 2021

DALL-E is an OpenAI model that can produce visuals based on textual descriptions. The public debut of this model took place in January 2021. DALL-E utilizes the GPT-3 model as its foundation to generate distinct and imaginative visuals according to the given textual input. DALL-E was introduced in 2021.

OpenAI created CLIP (Contrastive Language-Image Pre-training) and publicly announced it in January 2021. This model employs a contrastive approach to pre-training learning, enabling it to comprehend text and visuals concurrently. CLIP is specifically created to carry out various activities related to the correlation between text and images, such as image categorization, visual comprehension, and semantic examination. CLIP was formally introduced in January 2021. The next iteration of DeepMind's protein structure prediction engine, AlphaFold 2, was officially released in November 2020. AlphaFold 2 demonstrates exceptional precision in forecasting the three-dimensional configuration of proteins, marking a significant milestone in the realms of computational biology and structural biochemistry.

In 2021, InstructPix2Pix was introduced as a method for picture transformation using GANs. It enables users to submit instructions through textual images and observe the interactive results of the image transformation. An AI research team at Stanford University developed the system and presented it in a paper titled "InstructGAN: Instructable Generative Adversarial Network for Interactive Image Generation with Global Guidance" at the CVPR 2021 conference.

The Stable Diffusion XL, an expansion of the Stable Diffusion model created by OpenAI, was introduced in a research paper titled "Understanding and Enhancing Density Estimation in Diffusion Models" presented at the NeurIPS 2021 conference. The introduction of Stable Diffusion XL took place in 2021.

Github Copilot, a development tool developed jointly by GitHub and OpenAI, was publicly released in June 2021. Github Copilot employs the GPT-3 model to produce code and recommendations within the code editor, taking into account the surrounding environment and the requirements of the user.

11. In 2022

The introduction of DALL-E 2, an enhanced iteration of the DALL-E model created by OpenAI, took place in May 2022. DALL-E 2 enhances the functionalities of the original DALL-E by incorporating the capacity to generate superior quality and more lifelike images based on textual descriptions.

Imagene MidJourney, a mental health platform created by Imagene Labs, was introduced in 2022. The platform utilizes genetic analysis technologies to offer consumers a comprehensive understanding of their mental health, including potential hazards and specific remedies.

12. In 2023

ControlNet is an image processing technique that was developed by Google AI Research and the University of Cambridge.

13. In 2024

Google's Imagen 3 is a sophisticated text-to-image algorithm that generates more intricate and lifelike images with less visual imperfections compared to earlier iterations.

NVIDIA's EDM-2 is a highly efficient neural network structure specifically developed for generating images. It provides superior output quality and shorter training durations compared to earlier models.

The LATTE3D model developed by NVIDIA specializes in rapidly rendering text into 3D, delivering exceptional quality rendering of 3D objects with minimal delay. This is particularly advantageous for utilization in video games, creative endeavors, and virtual training settings.

C. Limitations of Generative AI

Despite the notable advancements in AI in recent years, it is essential to acknowledge the numerous limitations and deficiencies to explore the topic further. Some significant constraints are:

1. Dependence on quantity and quality of data [26]

Training generative models necessitates the use of a particular dataset, and the amount and caliber of this data have a direct influence on the resulting output. Insufficiently extensive and relevant data sets may lead to erroneous, inconsistent, or biased resultant products;

2. Difficulty with rare or new examples

Consequently, artificial intelligence may need help effectively handling certain circumstances not included in the training data. In such instances, this may yield impractical or inaccurate outcomes;

3. Limitations in original content creation [27]

Generative AI utilizes specific datasets to acquire knowledge and endeavors to replicate the patterns and regularities inside each dataset. It achieves this by amalgamating existing data and producing novel outputs. Nevertheless, this is the drawback. Artificial intelligence is limited in its ability to generate novel outcomes due to its dependence on existing data, resulting in diminished originality;

4. Interpretability and explanation

Several generative AI models, including deep neural networks, are sometimes called "black boxes," indicating that their processes and reasoning behind producing expected outputs are not understandable [27]. The absence of interpretation and comprehensibility is especially significant in domains where responsibility and openness are of utmost importance;

5. Lack of control

Generative models frequently need more precise manipulation of their output [28]. While it is possible to generate intricate content, it can be challenging to guide or shape the creative process to meet specific objectives or adhere to particular limitations;

6. Ethical issues and bias

Generative AI algorithms have the potential to magnify bias or uneven distribution of quantitative training data unintentionally [29], [30]. If the data exhibits bias, such as gender or racial bias, the final content may accurately reflect and include such biases. Addressing equity and mitigating prejudice in generative AI models continues to be a substantial obstacle;

7. Lack of context

Models frequently need more comprehension of the environment in which the results are generated, resulting in inconsistent outcomes [31]. They often generate text that is coherent but lacks evidence of profound comprehension or logical coherence;

8. Computing and time resources

The training and operation of AI models require significant computational resources, particularly high-performance equipment and substantial memory capacity [32]. Furthermore, the time needed for loading and processing frequently grows exponentially while the quality only improves linearly, raising concerns about the task's practicality. Computing units also consume a lot of electricity, leading to significant costs.

III. Problem Statement

This study seeks to objectively investigate the behavior of generative neural networks by employing various training methods and combinations of picture production tags. The objective is to examine the potential of this neural network as a "black box" for identifying concealed patterns that may remain imperceptible to humans, even when reviewing extensive datasets. This research holds significance and pertinence due to the remarkable performance of neural networks across various tasks, although the underlying reasons behind their decision-making processes remain unclear.

More precisely, this experiment investigated the subjects' behavior when exposed to various training techniques and image production labels. Individuals may get a further understanding of the process of making decisions. This study is significant as it enables us to harness the full potential of neural networks more efficiently and enhance their performance in diverse application domains. By comprehending the variables that affect the functioning of neural networks and examining the consequences of various training techniques and picture production labels, it is possible to progress in areas like computer vision, natural language processing, and creative content development [33].

Academics are highly focused on mitigating bias in different AI models [34]. Scientists recommend eliminating specific components of an object's descriptive attributes to prevent prejudice. Conversely, lacking essential attributes might result in a lack of objectivity. A viable resolution to this generative model quandary entails a lucid and comprehensive delineation of the characteristics of the item from which the image is generated. The primary objective of this research is to empirically identify the flaws of two image production methods by comparing them without delving into their specific qualities. In addition, you can utilize fundamental principles that can be tailored to your preferences and then evaluate the outcomes.

IV. Experiment Description

The basic concept uses images of a person in various positions and perspectives. The main idea of this experiment is to generate images from text prompts into images. The results of this custom concept will be compared with those generated images created based on specific prompts, and then

image generation will be explored with and without determining appropriate characteristic values. The experiment was carried out in several stages, the main of which were:

a. create a custom concept from a person's drawing with various perspectives and dimensions that are the same or close to the baseline;
b. training a baseline model with custom concepts;
c. creating images with different prompt tags using the model obtained in the previous stage;
d. Fréchet Inception Distance (FID) Score testing
e. analysis of results.

The paragraphs of this section describe each stage in more detail, and the next section of this document analyzes the results.

A. Defining Generative Models for Experiments

The study selected the Stable Diffusion generative neural model. Stable Diffusion is a freely available text-to-image model that may produce images using written descriptions. The mentioned model is a variation of the latent diffusion model, which falls under deep generative neural networks. This model can generate lifelike images by progressively eliminating random noise iteratively [35]. Stable Diffusion employs a variational autoencoder (VAE) to transform pictures into latent vectors that capture the image's high-level characteristics. Subsequently, it employs U-Net, a convolutional neural network with forward connections, to convert the latent vectors into a picture. U-Net also takes in a text encoder as input, which converts word descriptions into various vectors that govern the process of generating images. The text encoder employs the CLIP model, which is capable of capturing the semantic similarity between images and text. Stable Diffusion produces images by removing random noise and applying a series of smoothing procedures, each of which improves the realism of the image and gets it closer to a textual description. The model acquires the ability to identify images by undoing the action of introducing distortion to the training images. Therefore, the model can be trained on photos that contain both noise and are clean, resulting in a wide range of high-quality images. Stable Diffusion is a versatile technique that may be applied to various tasks, including generating images from scratch, altering existing images using text, completing missing details, and improving low-resolution photos.

B. Dataset Creation and Collection

Firstly, it is important to choose an appropriate dataset. Due to the inability to locate the necessary quantity of similar images online, the decision was made to create appropriate photographs using the Stable Diffusion technique. So, baseline data is created from pictures of people from various perspectives as a dataset with a custom concept. An example of a snapshot of the concept image can be seen in Figure 2.

Fig. 2. Custom concept with real photos from various perspectives as a baseline dataset

Fig. 3. Abnormal/deformed photos are used as a negative prompt.

In addition, a dataset of abnormal/defective photographs was also collected, which will be used as "negative prompts" to improve the resulting image quality. Figure 3 shows an example of an image snippet.

C. Model Training

Once the dataset is established, it is imperative to identify the fundamental model for subsequent training due to the numerous possibilities. The decision was made to select the Stable Diffusion model version 1.5, which was trained to generate images with a resolution of 512 x 512 pixels [36].

Stable Diffusion is an Artificial Intelligence (AI) generator technology that facilitates human activities in creating two-dimensional works of art. This technology allows users to generate images with simple text commands. Stable Diffusion is a generative AI technology capable of producing unique, realistic images based on text and image prompt commands (Figure 4). Stable diffusion is open source so that it can be used flexibly. This technology can train the system to produce a realistic picture of something the user describes.

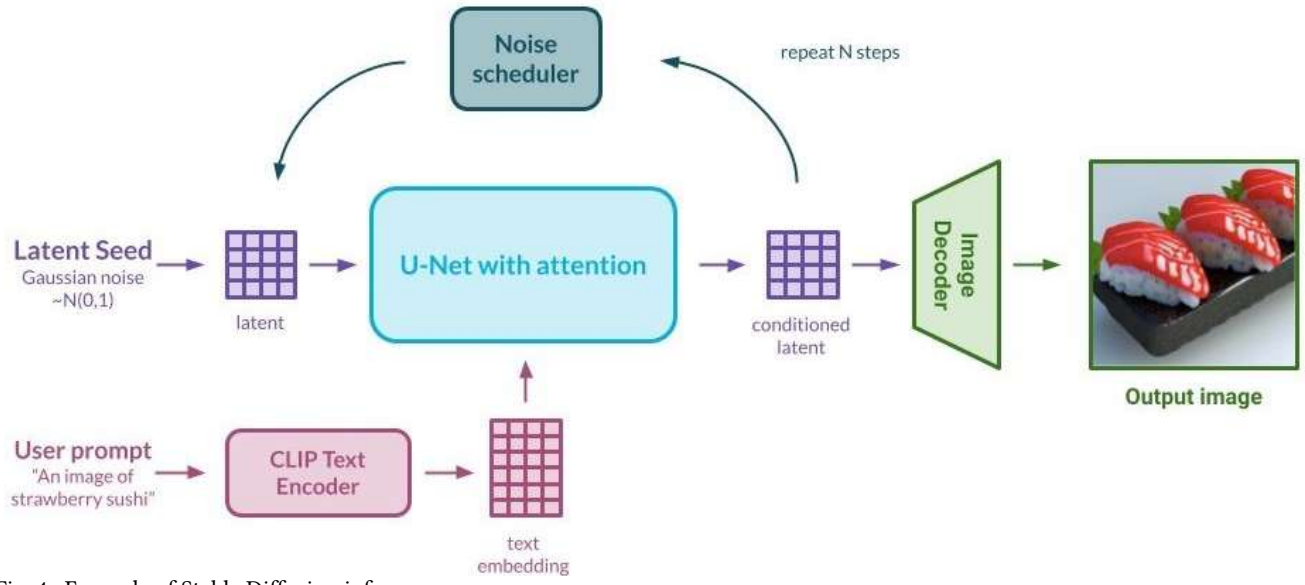

Fig. 4. Example of Stable Diffusion inference process

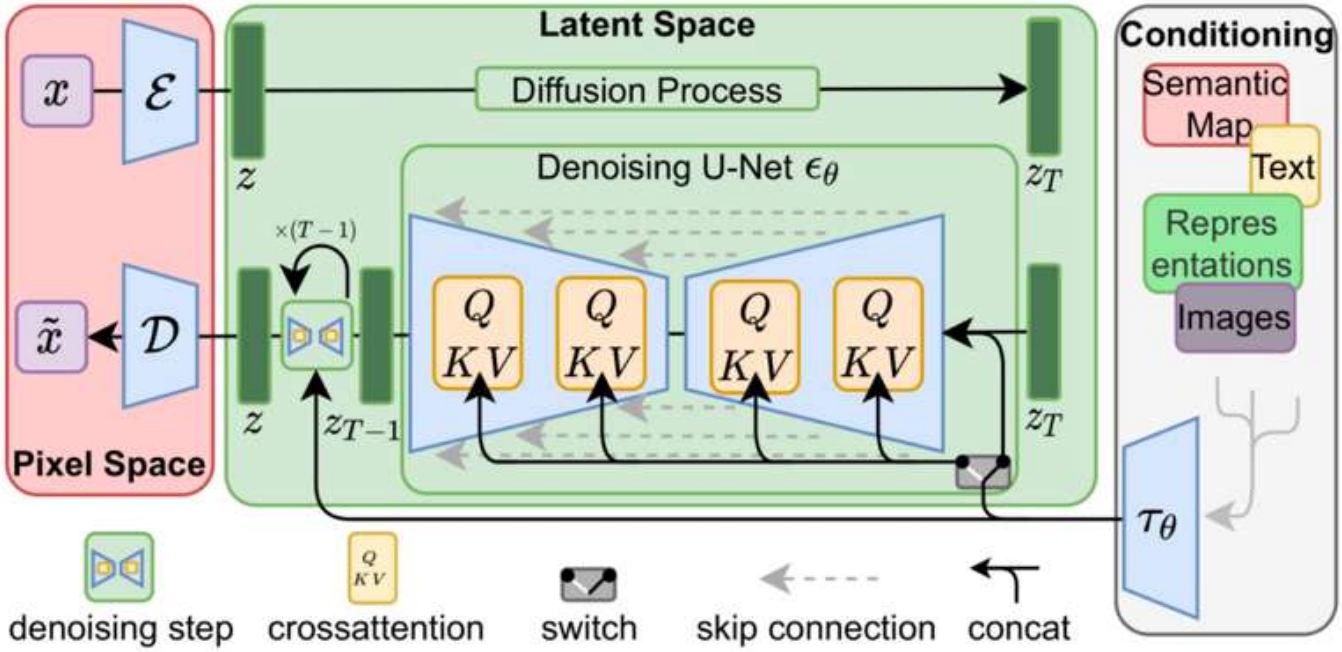

Fig. 5. Diagram of the latent diffusion architecture used by Stable Diffusion

The way Stable Diffusion works is different from many other image creation models. In principle, this AI generator uses Gaussian Noise technology to generate image code. After that, the system will use a noise predictor to recreate the image according to the commands. Stable Diffusion images or photo quality do not use pixels but low-volume latent space. The latent diffusion architecture diagram is shown in Figure 5. The aim is to speed up the image results that users need. Apart from that, Latent is considered to be able to produce more detailed and smooth images. With this system, Stable Diffusion can be used on devices with NVIDIA graphics cards and 8GB RAM. Stable Diffusion can produce images from various commands. They start from text to image, image to image, graphic artwork, image editing, and video creation.

Physics, especially thermal equilibrium, inspires the diffusion technique used to train AI. An example of thermal equilibrium is when ice melts with hot tea. Slowly but surely, all the tea water will have the same temperature. Another example is when gas fills the room, even though it moves randomly. In AI, diffusion techniques add noise to data -- such as images -- to change the data structure. However, unlike the diffusion process in chemistry - which cannot be reversed - AI trained using diffusion techniques can learn "reverse diffusion." As the name suggests, reverse diffusion is the process of recovering data whose structure has been changed by the diffusion process.

Significant memory capacity and a Graphics Processing Unit (GPU) are essential for model training, necessitating an external computational resource search. After evaluating the various alternatives, the decision was made to utilize the services provided by Google Colab.

Fig. 6. Generated Image Results Using Custom Concepts with the “Photo of a Person” Prompt

Neural networks possess a repository of familiar words (tokens) and their connections. Thus, avoiding using tokens for learning in frequently employed word forms is advisable, as doing so may result in a less accurate understanding of the topic. In this scenario, the network had prior knowledge of this idea, making it challenging to determine the exact fraction of the model that incorporates both old and new information during its development. Therefore, while instructing models about specific topics, it is advisable to utilize fictitious terms.

In this instance, it is essential to instruct the network on a comprehensive representation of the concept of a "photograph of an individual," with the instance prompt being "pakalam." Subsequently, the subsequent action involves loading all the photos into the runtime. Subsequently, it is essential to establish the parameters for the training process. Figure 6 displays the outcomes of the conducted training. Following the training process, many sets of photos were produced, each with distinct tags. These sets were used to investigate the variations among the trained models. The subsequent section of this study provides an analysis and presentation of the obtained data.

V. Result and Discussion

A. Generated Image

After creating the model, the resulting image is analyzed for compliance with the specified parameters. Verification starts with a model trained without descriptions/prompts. A collection of photos is generated with different commands to obtain images. Figure 7 shows pictures generated from several text-to-image prompts with custom concepts using the Stable Diffusion model.

The model successfully acquired the ability to generate visuals that closely resemble custom concepts and input data when prompted. Style, composition, shape, and scheme are according to the desired prompt input. The experiment consisted of several examples of input prompts plus negative prompts such as "bad anatomy, ugly, deformed, disfigured."

After analyzing the generation results, it can be seen that generative artificial intelligence (GenAI) succeeded in creating fake photos that are indistinguishable from real images. Even though we can see that there are deficiencies in Figure 7.k and Figure 7.l, there are still abnormal conditions in the hand. This aligns with the existing custom concept data as in Figure 2. The baseline image shows that the dominant object is the face; there are no hands. So, this affects the results of the generative artificial intelligence work with the stable diffusion model that has been carried out.

This finding was derived from the examination of a limited number of photographs. To conduct a generalizability assessment in this particular context, it is essential to undertake further investigations specifically tailored to investigate the offered hypotheses using statistical evaluation. Multiple iterations of increasingly precise instructions were executed to verify yield control. Overall, the trained model successfully handled the assignment. Despite some variances, the generated photos mainly adhere to the directions. However, in this instance, it is evident that the majority of the magnified facial images that have been produced validate the idea that was previously put forth.

B. Fréchet Inception Distance (FID) Score

The FID score measures the distance between the feature vectors of an actual image and the generated image. These feature vectors are extracted from a pre-trained Inception v3 network. A lower FID score indicates that the resulting image is more similar to the actual image, meaning the generative model produces higher quality and realistic output. For example, the Stable Diffusion paper mentions that architectural changes were introduced to improve the FID score of the diffusion model. Stable Diffusion has achieved impressive FID scores, allowing it to produce highly realistic images from text commands. So, in summary, when FID is mentioned in the context of Steady Diffusion and other image synthesis models, it refers to the Fréchet Initial Distance metric used to evaluate the realism and quality of the images produced by those models. The goal is to minimize the FID score between the generated and real images.

$$d_F(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\mu',\Sigma'))^2 = \|\mu-\mu'\|_2^2 + \mathrm{tr}\left(\Sigma+\Sigma'-2(\Sigma\Sigma')^{\frac{1}{2}}\right) \quad .. (1)$$

FID for images is defined in equation (1), in which the μ is the average value of the multidimensional distribution. The Image topic is the distribution of the activations of the last layer in the Inception V3 model from a set of real-world images. Correspondingly, the μ' is the activations from a generated image set. For example, it could be the images from the Stable Diffusion.

The test results with the FID score in the research experiment produced an FID Score value of 1284.4430. This shows that the score is still quite large, so the gap between the resulting and real images must be fixed.

(a) pakalam, (bearded:0.9) man with (slim:1.1) (fit:0.7) body, beautiful brown eyes wearing a nice italian navy suite, cinematic, 32k, clear focus,full body shot. hyperrealistic detailed character design concept art, matte painting

(b) pakalam, (bearded:0.9) man with (slim:1.1) (fit:0.7) body, stunning visuals, ultra quality, ultra detailed, ultra resolution, depth of field, masterpiece, highly detailed clothes, highly detailed body, perfect anatomy, symmetrical, (perfect face:1.2)

(c) pakalam, close up of a futuristic soldier, masterpiece, best quality, high quality

(d) Pakalam doesn't wear glasses, wears a neat suit with a cap, and has a beard and moustache

(e) pakalam doesn't wear glasses, has a beard and moustache

(f) Handsome actor pakalam very muscular strong massive russian prince in white imperial uniform at the imperial balcony , close-up, hyper detailed, trending on artstation, sharp focus, studio photo, intricate details, highly detailed

(g) Pakalam doesn't wear glasses, has a beard but no moustache

(h) pakalam, no mustache (bearded:0.9) man with (slim:1.1) (fit:0.7) body, stunning visuals, ultra quality, ultra detailed, ultra resolution, depth of field, ((stunning face:1.2))

(i) pakalam as the rock character from the movie The Royal Tenenbaums pastel colored background, in white, hyper - realistic photography, full body, 8k, close - up shot, close - up photo

(j) pakalam, smiled faintly, (bearded:0.1) man with (slim:1.1) (fit:0.7) body, beautiful brown eyes wearing an iron man suite, cinematic, 32k, clear focus,full body shot. hyperrealistic detailed character design concept art

(k) pakalam, wearing a hat, one hand visible, with a waterfall in the background.

(l) photo of pakalam full face, wearing dr.Strange suite

Fig. 7. The results generate custom text-to-image concepts with various input prompts using the stable diffusion model

VI. Conclusion

After analyzing the images produced by the model without descriptions/prompts, the results are as good and realistic as the original custom concept photos. After adding prompt input, the desired results were predominantly appropriate, and the generative artificial intelligence text-to-image process was successfully carried out, although there were still imperfections. This is in line with the original photo, which is dominant in the close-up view of the face, which influences the final result. The use of prompts that evoke other body parts could be better.

We hypothesize that the deficiencies arise from an ineffective approach to organizing descriptions or, more broadly, the impracticability of training the model on concepts beyond only face features, along with supplementary descriptions. Upon evaluating the generation results, it was determined that when a picture is generated without a specific prompt, and a portion outside the face becomes visible, it indicates the presence of an abnormal state or defect in that area. This conclusion is derived from examining a limited quantity of resultant photographs. To conduct a generalizability assessment in this context, it is essential to conduct more investigations specifically tailored to investigate the offered hypotheses using statistical evaluation.

However, training a model for a specific concept without additional descriptions and context may be more suitable for certain types of problems, as generative networks excel at integrating different prompts and nuances to create a comprehensive representation.

This AI model can help artists and designers create more attractive and realistic images and visualizations by creating highly realistic images from text descriptions. However, there are still several challenges to overcome, such as the need for quite a long computing time and quite a large amount of training data.

References

[1] Md. H. Al Banna et al., "Application of Artificial Intelligence in Predicting Earthquakes: State-of-the-Art and Future Challenges," IEEE Access, vol. 8, pp. 192880–192923, 2020 [Online]. DOI: 10.1109/ACCESS.2020.3029859

[2] A. Rai, "Explainable AI: from black box to glass box," Journal of the Academy of Marketing Science, vol. 48, no. 1, pp. 137–141, Jan. 2020 [Online]. DOI: 10.1007/s11747-019-00710-5

[3] C. Shorten, "Hierarchical Neural Architecture Search | by Connor Shorten | Towards Data Science," 2019. [Online]. Available: https://towardsdatascience.com/hierarchical-neural-architecture-search-aae6bbdc3624

[4] J. M. Durán and K. R. Jongsma, "Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI," Journal of Medical Ethics, vol. 47, p. medethics-2020-106820, Mar. 2021 [Online]. DOI: 10.1136/medethics-2020-106820

[5] M. Favaretto, E. De Clercq, and B. S. Elger, "Big Data and discrimination: perils, promises and solutions. A systematic review," Journal of Big Data, vol. 6, no. 1, p. 12, Dec. 2019 [Online]. DOI: 10.1186/s40537-019-0177-4

[6] B. Goertzel and C. Pennachin, Artificial General Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. DOI: 10.1007/978-3-540-68677-4

[7] V. C. Müller and N. Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," 2016, pp. 555–572. DOI: 10.1007/978-3-319-26485-1_33

[8] J. Clune, "AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence," 2020.

[9] R. Fjelland, "Why general artificial intelligence will not be realized," p. 1234567890, 2010 [Online]. DOI: 10.1057/s41599-020-0494-4

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015. DOI: 10.1038/nature14539

[11] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating Images from Captions with Attention," Nov. 2015 [Online]. Available: http://arxiv.org/abs/1511.02793

[12] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative Adversarial Text to Image Synthesis," International conference on machine learning, May 2016 [Online]. Available: http://arxiv.org/abs/1605.05396

[13] H. Zhang et al., "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," Proceedings of the IEEE international conference on computer vision, Dec. 2016 [Online]. Available: http://arxiv.org/abs/1612.03242

[14] T. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," Proceedings of the IEEE conference on computer vision and pattern recognition, Nov. 2017 [Online]. Available: http://arxiv.org/abs/1711.10485

[15] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, "Controllable Text-to-Image Generation," Advances in Neural Information Processing Systems, vol. 32, Sep. 2019 [Online]. Available: http://arxiv.org/abs/1909.07083

[16] A. Ramesh et al., "Zero-Shot Text-to-Image Generation," Proceedings of the 37th International Conference on Machine Learning, Feb. 2021 [Online]. Available: http://arxiv.org/abs/2102.12092

[17] M. Ding et al., "CogView: Mastering Text-to-Image Generation via Transformers," Advances in Neural Information Processing Systems, May 2021 [Online]. Available: http://arxiv.org/abs/2105.13290

[18] C. Wu et al., "N\"UWA: Visual Synthesis Pre-training for Neural visUal World creAtion," European Conference on Computer Vision, Nov. 2021 [Online]. Available: http://arxiv.org/abs/2111.12417

[19] J. Yu et al., "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation," Jun. 2022 [Online]. Available: http://arxiv.org/abs/2206.10789

[20] A. Nichol et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," Proceedings of the 38th International Conference on Machine Learning, Dec. 2021 [Online]. Available: http://arxiv.org/abs/2112.10741

[21] C. Saharia et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," May 2022 [Online]. Available: http://arxiv.org/abs/2205.11487

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Dec. 2021 [Online]. Available: http://arxiv.org/abs/2112.10752

[23] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," Apr. 2022 [Online]. Available: http://arxiv.org/abs/2204.06125

[24] C. Macdonald, D. Adeloye, A. Sheikh, and I. Rudan, "Can

ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis," Journal of Global Health, vol. 13, p. 01003, Feb. 2023 [Online]. DOI: 10.7189/jogh.13.01003
[25] "Generative AI – What is it and How Does it Work?," 2023. [Online]. Available: https://www.nvidia.com/en-us/glossary/generative-ai/
[26] "Data Dependencies | Machine Learning | Google for Developers," 2023. [Online]. Available: https://developers.google.com/machine-learning/crash-course/data-dependencies/video-lecture
[27] N. Ameen, G. D. Sharma, S. Tarba, A. Rao, and R. Chopra, "Toward advancing theory on creativity in marketing and artificial intelligence," Psychology & Marketing, vol. 39, no. 9, pp. 1802–1825, Sep. 2022 [Online]. DOI: 10.1002/mar.21699
[28] "What Are The Risks Of Google And Microsoft Advancing Their Generative AI Innovations?" [Online]. Available: https://www.forbes.com/sites/cindygordon/2023/05/11/google-strikes-back-on-microsoft/?sh=4d2a75ad463d
[29] D. Varona and J. L. Suárez, "Discrimination, Bias, Fairness, and Trustworthy AI," Applied Sciences, vol. 12, no. 12, p. 5826, Jun. 2022. DOI: 10.3390/app12125826
[30] X. Ferrer, T. van Nuenen, J. M. Such, M. Cote, and N. Criado, "Bias and Discrimination in AI: A Cross-Disciplinary Perspective," IEEE Technology and Society Magazine, vol. 40, no. 2, pp. 72–80, Jun. 2021. DOI: 10.1109/MTS.2021.3056293
[31] M. Kejriwal, "Artificial Intelligence needs to be more context-aware | by Mayank Kejriwal | Medium," 2021. [Online]. Available: https://mkejriwal1.medium.com/artificial-intelligence-needs-to-be-more-context-aware-ecc3097ee2ea
[32] "AI is harming our planet: addressing AI's staggering energy cost," 2022. [Online]. Available: https://www.numenta.com/blog/2022/05/24/ai-is-harming-our-planet/
[33] A. Bozkurt, "Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift Introduction: Generative AI and the next big thing (!)," Asian Journal of Distance Education, vol. 18, no. 1, 2023 [Online]. DOI: 10.5281/zenodo.7716416
[34] L. Moerel, "Algorithms can reduce discrimination, but only with proper data," 2018. [Online]. Available: https://iapp.org/news/a/algorithms-can-reduce-discrimination-but-only-with-proper-data/
[35] "Stable Diffusion Public Release — Stability AI," 2022. [Online]. Available: https://stability.ai/news/stable-diffusion-public-release
[36] "runwayml/stable-diffusion-v1-5 · Hugging Face." [Online]. Available: https://huggingface.co/runwayml/stable-diffusion-v1-5

Alam Rahmatulloh

is a lecturer and researcher in the field of Informatics at the Department of Informatics, Faculty of Engineering, Siliwangi University. He obtained his Master's degree in Informatics from the School of Electrical and Informatics Engineering, Bandung Institute of Technology 2015. His research interests include Microservices, Artificial Intelligence, IoT, and Cryptography. He is experienced in making information systems such as smart campuses and academic systems.