Vol.12 -1 (2025) 43–66
International Journal of Computers and Information
(IJCI)
Available online at *https://ijci.journals.ekb.eg/*

Animating Text Descriptions into Characters: A Comparative Review of Generative Models

Sameh Zarif [a], Khalid Amin [b], Abdalfatah Najjar [c], Marian Wagdy [d]

a Information Technology Department Menofia University, Egyptian Russian University, sameh.shenoda@ci.menofia.edu.eg, sameh-zarief@eru.edu.eg

b Information Technology Department Menofia University Menofia, EGYPT, kh.amin.0.0@gmail.com

c Information Technology Department Menofia University Menofia, EGYPT anajjar@ppu.edu

d Information Technology Department Tanta University Tanta, EGYPT, marian_wagdy@ics.tanta.edu.eg

Abstract

In recent times, the advent of text-to-image generative AI technologies, commonly referred to as AI Image Generators, has captured widespread attention for their remarkable capability to swiftly produce visuals based on textual descriptions. This development has ignited diverse perspectives, especially within the animation sector, making it a focal point for scholarly investigation due to the introduction of generative adversarial networks. Despite the progress, the domain confronts several challenges that necessitate further scholarly inquiry, including the production of high-resolution imagery featuring multiple elements and the creation of evaluation metrics that align with human assessment. Moreover, the outcomes of this study reveal that AI Image Generators hold the potential to substantially boost creative outputs in animation by providing a variety of characters and settings with superior visual quality. This review aims to examine and compare the extensive body of work in this field comprehensively. It will detail the algorithms employed, identify existing issues, and propose new directions for research.

Keywords: Generative-art; AnimeGAN; text-to-image; animation-character; generative-adversarial-network.

1. Introduction

The advancement of deep learning and AI technologies in recent years has revolutionized the animation industry by significantly reducing the time required to animate objects. These innovations have broadened the horizons of animation, enabling the creation of more sophisticated animations with less effort compared to traditional manual techniques. As a result, the once labor-intensive animation process has become more efficient, allowing animators to devote their energies to more inventive aspects that enhance the quality of animations [1,2,3,4,5].

Intelligent systems are now producing images for a wide array of applications, spanning education, entertainment, and providing accessible solutions for individuals with physical disabilities. Digital graphic design tools are becoming indispensable in the creation and editing of contemporary culture, art, and extending their benefits to other domains such as visual educational content, gaming, and animation. Notably, OpenAI's introduction of DALL-E, an AI Image Generator, has significantly impressed the public by its ability to generate images from textual prompts in seconds. According to OpenAI, between April and September 2022, DALL-E attracted 1.5 million users and managed to produce an average of 2 million images a day [6].

The animation sector values the precision with which AI can render characters, employing neural phase functions to achieve appearances that are more realistic and natural than those produced by traditional techniques, often exceeding user expectations in terms of speed.

Categorizing characters in animated films into distinct archetypes is a widely recognized and utilized practice among animators globally. Viewers, through experience, have developed expectations about character roles; for instance, a character depicted as a quarrelsome and irritating child is anticipated to also possess endearing qualities, while elderly characters are expected to be expressive and articulate, and antagonistic characters to exhibit cruelty and ruthlessness. Consequently, the design of animated characters incorporates three primary considerations: personality traits, which influence how a character interacts with others; physical characteristics, which affect movement and how characters are perceived and interpreted by viewers; and the physical condition or status, which shapes a character's responses and behavior across different scenarios. An example provided in Fig.1 illustrates a character with a short stature, and shortened limbs, highlighting how these aspects contribute to character development.

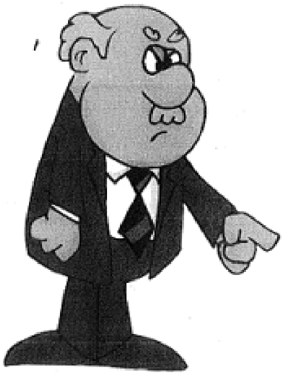


Fig.1 Animation character for short Guy [2].

This comparative survey aims to encapsulate and contextualize advancements in generative text-to-image models over the last seven years, providing an overview of the current research landscape and identifying areas that warrant further investigation. It critically examines the prevailing methodologies for evaluating text-to-image models, pinpointing notable deficiencies in existing metrics. The survey advocates for the development of superior datasets and evaluation standards, alongside enhancements in model training and architectural frameworks, as pivotal future research directions for generating animation character models from textual descriptions. This study exclusively focuses on the progression and assessment of methods for converting text-based descriptions into animation character models. It surpasses existing surveys in this domain by incorporating a broader spectrum of approaches, conducting an in-depth analysis of current evaluation techniques, and systematically addressing outstanding challenges in the field of text-based description to animation character generation [3,4,5,7,8,9].

This paper opens with an introduction that establishes the context for generating facial images from text descriptions, underscoring the significance and potential uses of this area of study. The section on fundamentals offers a broad overview of the essential concepts and methods used in creating facial images directly from text, covering the various algorithms and techniques essential for this task. Progressing, the paper reviews key models employed in generating animated characters from textual descriptions, with a particular focus on Conditional Generative Adversarial Networks (cGANs). It then compares the outcomes produced by each model, taking into account factors like the quality of the images, their diversity, and the computational resources required. Additionally, it evaluates the metrics used to measure the performance of these models and makes suggestions for future improvements in these evaluation methods. The section on challenges discusses the obstacles and complexities faced in generating facial images from text, including issues related to ambiguity and the understanding of meaning. In conclusion, the paper summarizes its main findings, reiterates the potential of this research area, and outlines directions for further investigation.

2. BASICS

The following sections of this paper are dedicated to exploring methodologies for converting text-based descriptions into animated characters, highlighting four key concepts vital for understanding the process [10]. In recent years, text-to-image generative artificial intelligence (GAI) technologies have achieved significant advancements [11]. The beginning of 2021 was particularly notable in AI research, marking the introduction of a technology referred to as the AI Image Generator. This breakthrough is credited to advancements in text-to-image synthesis, utilizing transformer techniques to achieve superior outcomes in terms of image quality, the relevance of text-to-image correspondence, and comprehensive domain analysis [12].

Firstly, this paper investigates the foundational method known as the unconditional Generative Adversarial Network (GAN), which utilizes random noise as input to produce images. This is succeeded by an analysis of the conditional GAN (cGAN), a variation that allows the image generation process to be conditioned on a specific label or attribute. Following this, the discussion shifts to the use of text encoders, which are employed to create embeddings from textual descriptions that act as conditioning factors for image generation. Lastly, the paper reviews the datasets commonly used in this area of research, highlighting their role in training and evaluating the performance of generative models.

2.1 Generative Adversarial Networks (GAN)

The original Generative Adversarial Networks (GANs) developed in is like two players competing in a two-player mini-max game, which consists of a generator G and a discriminator D: The generator tries to deceive the discriminator while the discriminator tries to discriminate between real training data and fake images. On V (D, G), D and G specifically engage in the following game as in Eq.1 [10]:

$$\min_{G}\max_{D} V(D,G)=\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)]+\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

According to Goodfellow et al., [10], this mini-max game has a global optimum precisely when pg $\sim$ pdata, and it converges to pdata under simple circumstances (such as when G and D have sufficient capacity). In reality, D's initial training samples are of very low quality and are confidently rejected by D where pdata is the real data distribution and pg is the noise distribution of Gaussian. It has been discovered that maximizing log(D(G(z)) rather than minimizing log (1 - D(G(z)) works better in practice.

2.2 Conditional Generative Adversarial Networks(cGAN)

Gaining control over the image generation process has significant practical value, despite the fact that the process of creating new, realistic samples is intriguing. Mirza et al. proposed the conditional GAN (cGAN) by specifying which MNIST[1] digit to produce with a conditioning variable y (such as class labels) at both the generator and discriminator. A single hidden layer Multi-Layer Perception (MLP) network combines images and labels for the discriminator [13,14]. Eq.2.

$$\min_{G}\max_{D} V(D,G)=\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x\mid y)]+\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z\mid y)))] \quad (2)$$

The cGAN objective function was extended in a number of variants to enhance conditional GAN training. For instance, the authors of AC-GAN suggested incorporating an auxiliary classification loss into the discriminator, which is denoted by the symbol LC [15]. As in Fig.2.

[1] MNIST is a widely used dataset of handwritten digits that contains 60,000 handwritten digits for training a machine learning model and 10,000 handwritten digits for testing the model.

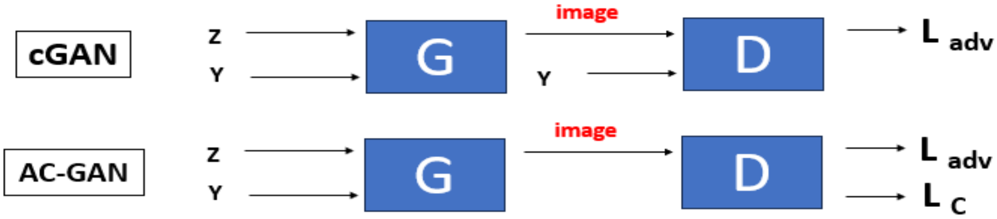


Fig.2 Simplified cGAN and AC-GAN architectures.

2.3 Text Encoding Techniques

It is not easy to make an embedding from textual representations that the network can use as a conditioning variable. Reed et al. [16] employ a pre-trained character-level convolutional recurrent neural network (char-CNN-RNN) to obtain the text encoding of a textual description. Traditional text representations like Word2Vec [17] and Bag-of-Words [18] were also shown to be less effective. Were TAC-GAN methods [19,20] is utilized by using Skip-Thought vectors[2].

The creators of StackGAN [21] introduced Conditioning Augmentation (CA) to randomly sample the latent variable from a Gaussian distribution as opposed to using the fixed text embedding achieved by a pre-trained text encoder. During training, a regularization term is utilized based on the Kullback-Leibler (KL) divergence between a standard Gaussian distribution and the conditioning Gaussian distribution. The bi-directional LSTM (BiLSTM) [22] was used by the developers of AttnGAN [23] to replace the char-CNNRNN in order to extract feature vectors by concatenating the hidden states of the BiLSTM to create a feature matrix for each word.

2.4 Top Used Datasets for Text to Image Synthesis Models

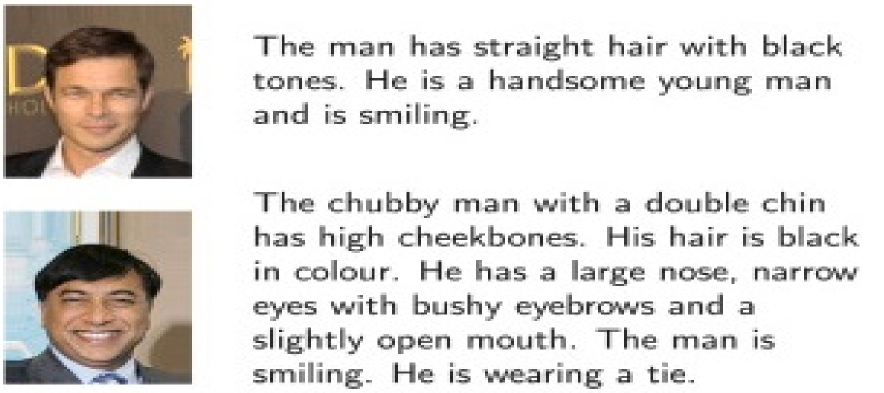
The foundation of any machine learning challenge is a dataset. The datasets Oxford-102 Flowers [24], CUB, CUB-200 Birds [25], and COCO [26] are often used in text-based description to animation character research. Both the CUB-200 Birds and Oxford-102 Flowers collections have about 10,000 photos each. Each image portrays a single object and there are 10 corresponding subtitles per image. On the other hand, COCO has around 123k photos with five captions each. Images in the COCO dataset typically features many, frequently interacting items in complicated backgrounds, unlike Oxford-102 Flowers and CUB-200 Birds. The COCO is used in the majority of text-based description to animation character studies. Recently the Large-scale Artificial Intelligence Open Network (LAOIN-5B) [27] dataset which is built by 14x larger than its predecessor LAOIN-400M. LAOIN-5B is one of the largest image-text datasets and it is available free for everyone.

3. Generation-only-Facial Image from Text

The idea of producing images from a written description started to take a different turn as artificial intelligence research, particularly in the area of deep learning. The first idea was to produce only a facial image from words. An example of these methods is the FTGAN model which is a Fully-trained Generative Adversarial Network for Text to Face Generation [28].

This study suggested a unique text-to-image network called FTGAN that simultaneously trains the text-encoder and image-decoder. FTGAN demonstrates its dominance over the most recent state-of-the-art network through trials using the open-source dataset CUB where the common datasets used in this field is CUB, Oxford102, and COCO. An example is shown in Fig.3.

[2] Skip-Thought Vectors” or simply “Skip-Thoughts” is the name given to a simple Neural Networks model for learning fixed-length representations of sentences in any Natural Language without any labeled data or supervised learning



The man has straight hair with black tones. He is a handsome young man and is smiling.

The chubby man with a double chin has high cheekbones. His hair is black in colour. He has a large nose, narrow eyes with bushy eyebrows and a slightly open mouth. The man is smiling. He is wearing a tie.

Fig.3 An example of text-to-face synthesis.

Another model is called STYLET2F which generates human faces from textual description using STYLEGAN2 [29]. In this paper, a comprehensive method for generating human faces from textual descriptions was presented. This method was built by utilizing StyleGAN2's power [30] and subsequent research into its latent space. In addition to controlling a set of facial features that adequately describe a human face, this approach provides a consistent mapping between the input text and the generated images as shown in Fig.4.

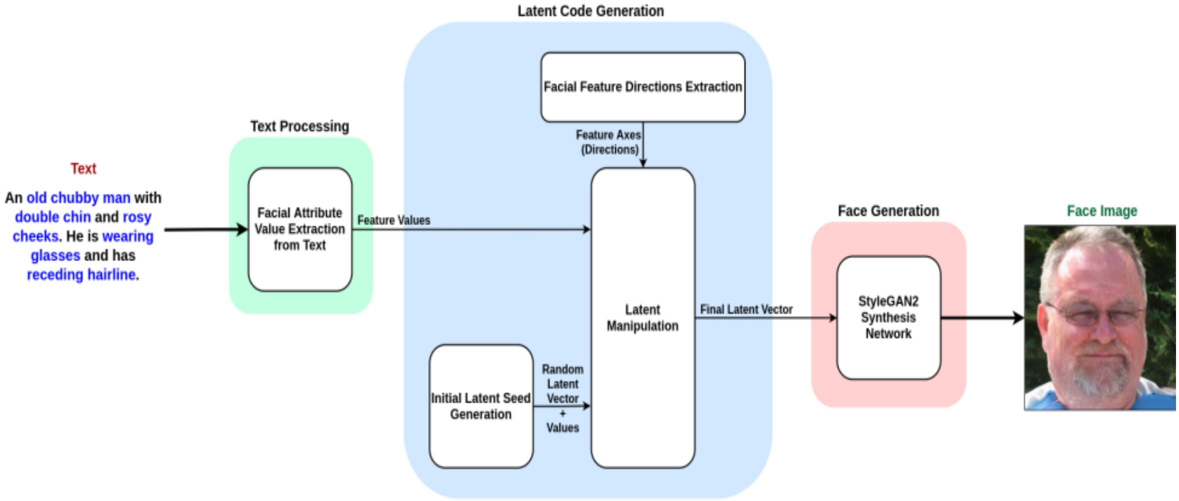

Latent Code Generation
Facial Feature Directions Extraction
Feature Axes (Directions)
Text Processing
Text
An old chubby man with double chin and rosy cheeks. He is wearing glasses and has receding hairline.
Facial Attribute Value Extraction from Text
Feature Values
Latent Manipulation
Final Latent Vector
Face Generation
StyleGAN2 Synthesis Network
Face Image
Initial Latent Seed Generation
Random Latent Vector + Values

Fig.4 Face generation from text system overview. The system consists of 3 stages. First, the input text is processed to extract the facial attribute values. Then, these feature values are used to manipulation StyleGAN2 latent space, in order to sample the latent code that represents such features. Finally, the extracted latent code is passed to StyleGAN2 synthesis network to generate the final face image [29].

Efforts are still continuing to produce only a face image through text, as the ManiCLIP model which is a Multi-Attribute Face Manipulation from Text tackles the multi-attribute face editing problem with text input. To achieve minimal excessive editing, they introduce a decoupling training scheme and add entropy loss to the learning objective [31]. As a result, the edited images not only match the given text but also are natural and have minimal irrelevant attribute change, maximally preserving the original identity as shown in Fig.5.

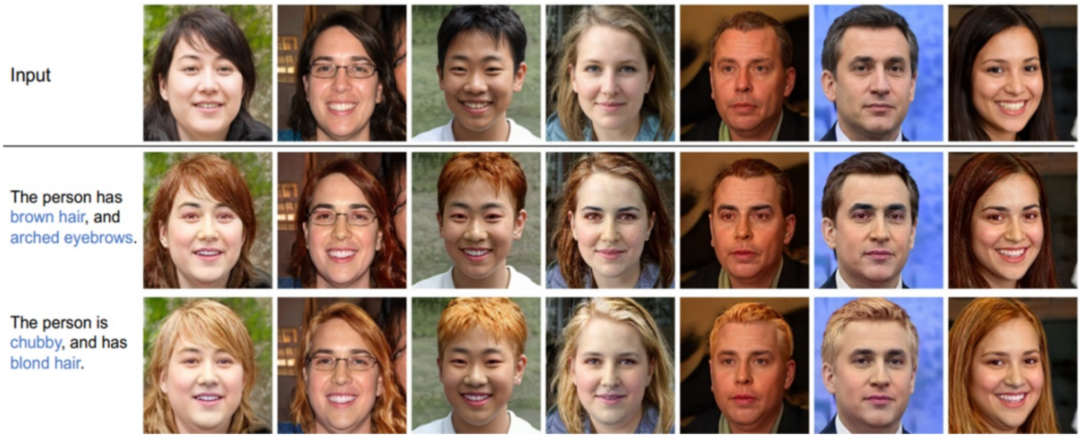

Input
The person has brown hair, and arched eyebrows.
The person is chubby, and has blond hair.

Fig.5 ManiClip model examples. This model is able to naturally edit multiple face attributes from natural language instructions. Top row shows the original images. Rows 2-3 show the edited face images under different text descriptions [31].

Recent advancements have introduced a novel model titled "Text-Free Learning of a Natural Language Interface for Pre-trained Face Generators." This innovative approach trains a pre-existing image generator to interact with pre-trained text embeddings (CLIP) in conjunction with StyleGAN [32]. The model's interface facilitates rapid conditional image synthesis directly from natural language inputs following the training phase. Remarkably, this method requires only an unlabeled set of images for training, eliminating the need for any associated textual data. Furthermore, the introduction of Conditional Variational Autoencoders (CVAEs) extends the model's versatility, allowing it to generate diverse visual features not explicitly detailed by the text prompt. The adoption of a nonparametric, sampling-based testing method significantly enhances the generation's quality and robustness. This is achieved by modelling a non-parametric distribution within the space of text embeddings as represented in CLIP's image embeddings. Empirical evidence suggests that the Fast text2StyleGAN is capable of producing high-quality images that more accurately reflect the nuances of the text prompts.

4. Common Models that Generate Animation Characters from Text-Based Description

Model 1: Generative Adversarial Text to Image Synthesis

In this study, a straightforward and efficient approach for creating images from thorough descriptions was created. The researcher showed how the model could combine a variety of realistic visual interpretations of a given text caption. The text-based description of animation character synthesis on CUB was significantly enhanced by this manifold interpolation regularization. They demonstrated how to separate style and content, as well as how to move bird poses and backgrounds from query images onto text descriptions. Finally, their findings from the MS-COCO dataset show how generalizable this method is for creating images with multiple objects and different backgrounds [33].

Model 2: StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

The architecture of STACKGAN is a two-stage generative adversarial network (GAN) that aims to generate highly realistic images from text descriptions. In the first stage, known as Stage I, a text encoder and a generator work together to produce low-resolution images conditioned on the input text. This stage allows the network to capture the global structure and overall appearance of the desired image. The generated low-resolution images are then passed to the second stage, Stage II, where a conditioning mechanism based on the text description is applied. The Stage-II generator takes these conditioned images and refines them to a higher resolution, producing more detailed and visually appealing results. By incorporating the conditioning mechanism in both stages, STACKGAN effectively aligns the generated images with the input text, resulting in coherent and contextually accurate visual representations. The hierarchical structure of STACKGAN enables the generation of high-quality images that exhibit both global coherence and fine-grained details, offering a promising approach for text-to-image synthesis tasks. The dataset mentioned in the paper is the CUB (Caltech-UCSD Birds-200-2011) dataset with Inception Score (IS) for the CUB dataset is reported as 3.70 ± 0.04 [34]. The suggested approach is shown in Fig.6. This method creates images with a higher resolution (e.g., 256X256) and more photo-realistic details and diversity when compared to existing text-to-image generative models.

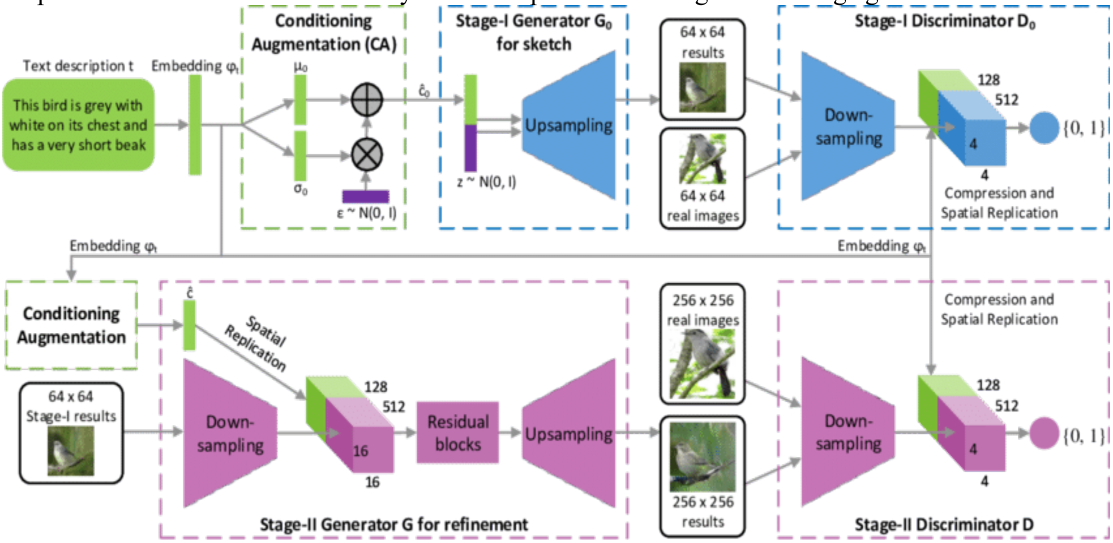


Fig.6 The architecture of the proposed StackGAN. The Stage-I generator draws a low-resolution image. The Stage-II generator generates a high-resolution image with photo-realistic details [34].

Model 3: Hierarchical Text-Conditional Image Generation with CLIP Latents

It has been demonstrated that robust image representations that capture both style and semantics can be learned using contrastive models like CLIP (Contrastive Language–Image Pre-training). The paper proposes a two-stage model for image generation that makes use of these representations. The first stage is a prior that generates a CLIP image embedding. The second stage is a decoder that generates an image as shown in Fig.7. The model demonstrates that without sacrificing photo-realism or caption similarity, explicitly generating image representations increases the diversity of images. The decoders adapted to picture portrayals can likewise create varieties of a picture that save the two its semantics and style while shifting the trivial subtleties missing from the picture portrayal.

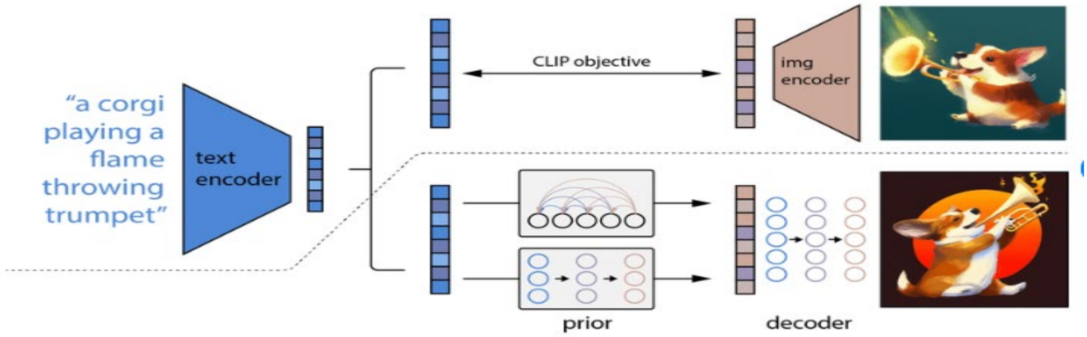
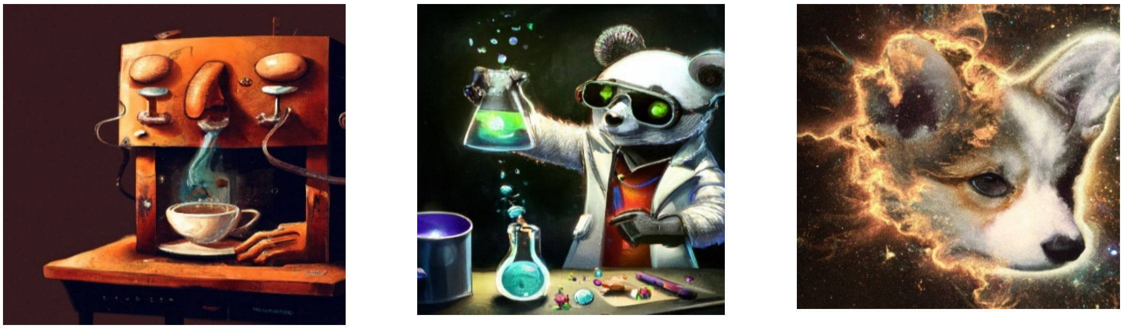


Fig.7 A high-level overview of CLIP. Above the dotted line, CLIP training process. Below the dotted line, text-to-image generation process [35].

Additionally, Clip's joint embedding space makes zero-shot language-guided image manipulations possible. For the decoder, they use diffusion models, and for the prior, they try autoregressive and diffusion models [35]. They find that diffusion models are more computationally efficient and produce better samples as shown in Fig.8.



an espresso machine that makes coffee from human souls, art station

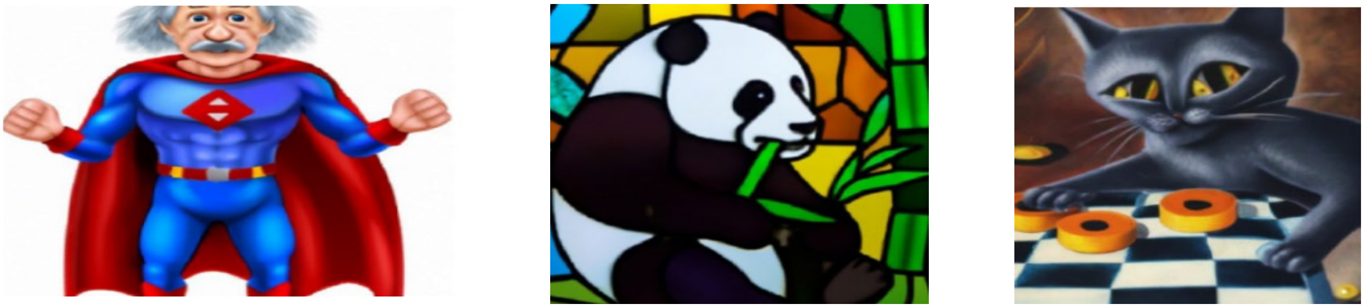
panda mad scientist mixing sparkling chemicals, art station

a corgi's head depicted as an explosion of a nebula

Fig.8 Samples from a production version from CLIP Latent model

Model 4: GLIDE: Towards Photo-realistic Image Generation and Editing with Text-Guided Diffusion Models

Guided Language-to-Image Diffusion for Generation and Editing (GLIDE) can be used to edit images by using natural language also can prompt text to insert new objects, create shadows, create reflections. and implement images in painting. It can also convert basic line drawings into realistic ones with exceptional capabilities for manufacturing and repairing. Recent research has shown that probability-based scrutiny models can also produce high-quality synthetic images, particularly when combined with diversity and sincerity as shown in Fig.9. This figure displays some samples generated from the model. The dataset used is MS-COCO with FID score reported in the paper for their model is 12.24 and the Inception Score (IS) for the MS-COCO dataset is approximately 19. [36].



an illustration of albert Einstein wearing a perhero costume

a stained-glass window of a panda eating bamboo

a surrealist dream-like oil painting by Salvador of a cat playing checkers

Fig.9 Selected samples from GLIDE using classifier-free guidance [33]

OpenAI[3] nitrated a directed diffusion model in May-2022, which allows diffusion models to be conditional on classifier labels. GLIDE works to improve this success by providing a guided deployment of a text-modal image generation problem. For natural language descriptions, researchers tested two guiding strategies after

[3] It is an artificial intelligence (AI) research laboratory consisting of the for-profit corporation https://openai.com

training a (GLIDE 3.5 billion-variable) diffusion model using a text-to-sharp encoder. Two alternatives are CLIP instructions and workbook-free instructions. CLIP is a scalable technology for learning shared representations of text and images that deliver a score based on how close the image is to the caption. The team used these as their own diffusion models by replacing the classifier with a CLIP 'guiding' model. At the same time, CLIP is a routing strategy for diffusion slaughterhouses that does not involve separate cleaver training.

The GLIDE architecture is structured around three primary components. The first component is an Ablated Diffusion Model (ADM) tasked with generating images at a resolution of 64x64 pixels. Following this, a text conversion model integrates textual descriptions into the image generation process through a text vector. The final component is a down-sampling model, which escalates the resolution of these micro-images from 64x64 to a more discernible 256x256 pixels. The synergy between the first two components ensures the generated images accurately reflect the intended text vector, while the third component enhances image clarity for better interpretation.

GLIDE incorporates an attention-facilitating conversion mechanism, offering nuanced control over image output by processing textual input prompts. This is achieved by training an adapter model on a comprehensive dataset of images and annotations, akin to the methodology employed by Project DALL-E from OpenAl. Textual inputs are initially encoded into a sequence of K symbols, which are then transformed. The output from this converter can interact with the ADM model in two distinct ways: replacing the class token or independently rendering the final layer of token embedding across the dimensions of each embedding layer in the ADM model. This innovative approach allows the ADM model to create images from new combinations of text symbols in a realistic manner, grounded in the learned correlations between words and corresponding images. The text conversion component boasts 1.2 billion parameters and utilizes an expansive architecture with 24 layers and a width of 2048. The advanced model, equipped with approximately 1.5 billion parameters, distinguishes itself from the foundational model by featuring a more capable text encoder with enhanced channel capacities.

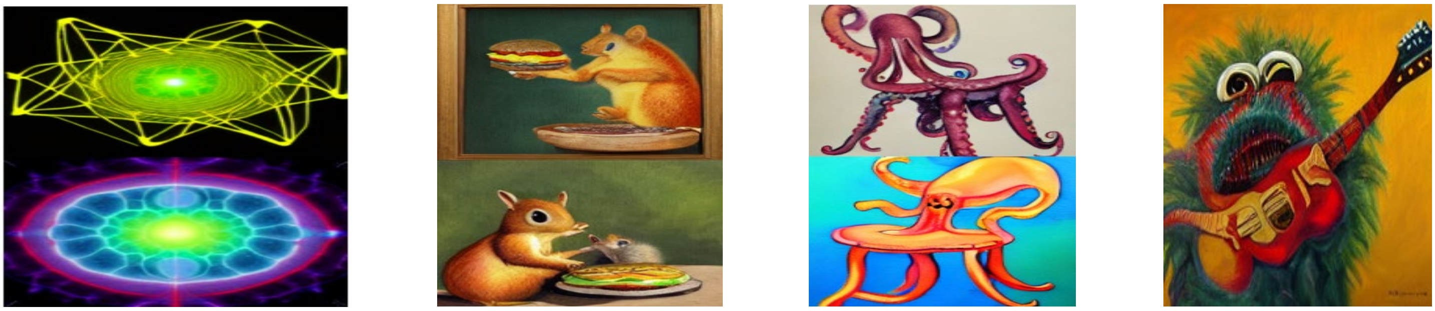
This development underscores the remarkable capability of computers not just to describe images, but to generate them from textual descriptions, marking a significant departure from traditional image search methods like Google's, which merely retrieve existing images. OpenAI has been at the forefront of this innovation, achieving notable success with their GLIDE model. Trained on hundreds of millions of images, GLIDE demonstrates superior image realism when compared to its predecessor, DALL-E, illustrating the profound potential of generative AI in creating images from text descriptions.

Model 5: High-Resolution Image Synthesis with Latent Diffusion Models

Diffusion models (DMs) have emerged as forefront technologies in the synthesis of image data and beyond by incrementally constructing an image through a series of de-noising steps. These models are particularly noted for their ability to allow for controlled guidance of the image generation process without necessitating retraining. However, the primary limitation of DMs stems from their operation within the pixel space, which results in substantial computational demands—often requiring hundreds of GPU days for optimization and incurring high costs during inference due to the need for sequential processing. The dataset used is MS-COCO with FID score for text-conditional image synthesis on the MS-COCO dataset is reported as 12.24. and the Inception Score (IS) for the same is 20.03 ± 0.33 [37].

To circumvent these constraints, the model leverages the latent space of powerful pre-trained auto-encoders, facilitating the training of DMs with reduced computational resources without compromising on their quality and flexibility. This approach marks a significant advancement by achieving an optimal compromise between simplifying complexity and preserving detail, thereby substantially improving visual fidelity.

By integrating cross-attention layers within the architecture, diffusion models are transformed into versatile generators capable of handling a wide range of conditioning inputs, such as text descriptions or bounding boxes, effectively broadening their applicability. This inclusion allows for high-resolution image synthesis through a convolutional method. Compared to traditional pixel-based diffusion models, Latent Diffusion Models (LDMs) set new benchmarks in performance across several metrics, excelling in tasks like image in-painting, class-conditional image synthesis, and demonstrating highly competitive outcomes in text-to-image synthesis, unconditional image generation, and super-resolution tasks. Fig.10 exemplify the high-quality images generated by LDMs, showcasing their superior capabilities and the broad spectrum of their application potential.



An illustration of a slightly conscious neural network

A painting of a squirrel eating a burger

A watercolor painting of a chair that looks like an octopus

Animation character (monster carrying a guitar)

Fig.10 Samples for user-defined text prompts from the model for text-to-image synthesis, LDM-8 (KL), which was trained on the LAION database [37].

Model 6: Stable diffusion model v1

The Stable Diffusion Model v1 stands as a probabilistic, generative model that capitalizes on diffusion processes to craft high-quality images. Through a sequence of iterative refinements starting from a noise sample, it meticulously shapes coherent and lifelike images. Central to its mechanism is a diffusion process that incrementally introduces noise via a series of learnable transformations, thereby approximating the data distribution with high fidelity. When these transformations are applied in reverse, beginning with a generic image state, the model adeptly sharpens and enhances the visual output to produce images that are both sharper and more aesthetically pleasing.

One of the hallmarks of the Stable Diffusion Model v1 is its robust training stability and the remarkable efficacy it displays in image synthesis tasks, positioning it as a significant advancement in generative modelling. Notably, this model has achieved groundbreaking performance, marked by an FID (Frechet Inception Distance) score of 7.27 on the COCO dataset, a feat accomplished without direct training on COCO data [38].

Illustrated in Fig. 11, the model's prowess is exemplified through various outputs, underlining two principal strengths. Firstly, it demonstrates an exceptional capacity to generate images that accurately reflect provided textual descriptions. Secondly, it showcases an innate ability to produce imaginative and creative visuals from mere images fed into the neural network. This dual capability underscores the model's versatility and its potential to revolutionize the generation of digital imagery, bridging the gap between descriptive inputs and visual creativity.

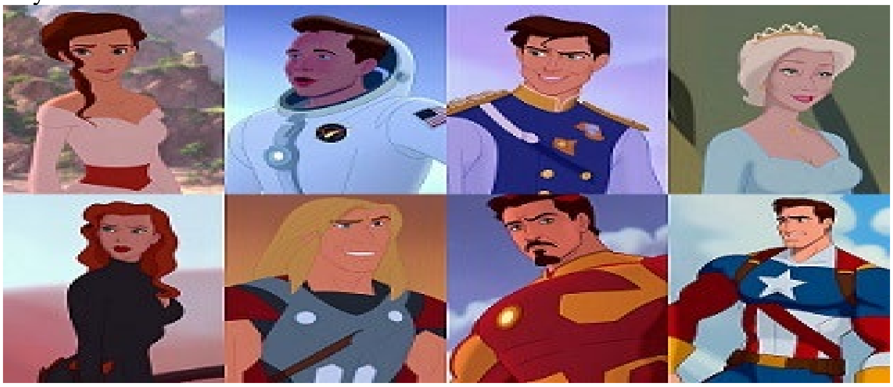


Fig.11 Animation Characters rendered with Stable diffusion model.

Another simple example is if we give the proposed model the description of "a dog in a black hat," the result will be as shown in Fig.12.

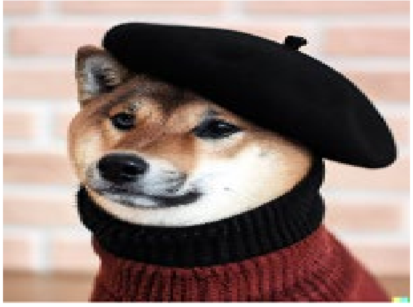


Fig.12 A dog in a black hat rendered with Stable diffusion model.

Now let's take an example of the second point. Suppose that we were given the proposed model a picture of a crater, The result will be the same image creatively.

This system comprises two stages: the initial stage involves training the CLIP (Contrastive Language-Image Pre-training) model, while the second stage employs the Diffusion model as the decoder. The first step of the model focuses on training a distinct model that aims to identify and extract the essential features that connect to the bulk image. For instance, when provided with a text input such as "Draw a picture of a dog," this model will learn to identify the crucial characteristics of a dog image, such as the eyes and the defining lines that constitute the dog's visual representation.

This means that a large database has been built that contains millions of sentences and returns their own pictures. So, the output of the first model is the characteristics of the images with the characteristics of the sentences that are related to the image. It means each image has its own characteristics that correspond to it in the sentence. Fig.13 shows an example when we prompt the text (astronaut on a horse).

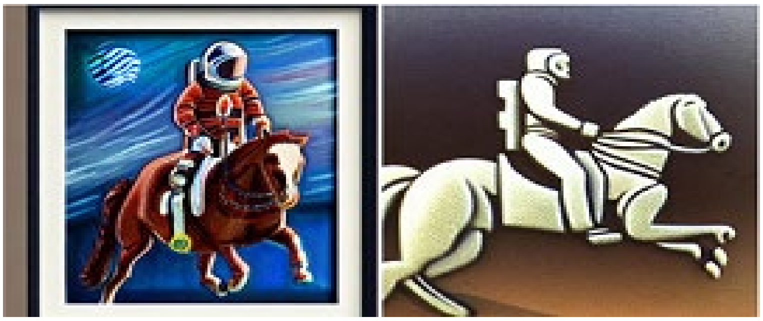


Fig.13 An image of astronaut on a horse rendered with Stable diffusion model.

The idea of the Diffusion Model is that there is a set of images then random numbers of Gaussian noise are added to it until the image becomes blurred. Then the process is reversed until the real image appears. This is using time which means every time the process is done at ti. This method has proven its effectiveness. Here, the Diffusion Model is used in a different way. The input will be the image embedding's properties, which were the output of the prior. Some experiments in the following figures like Fig.14 that shows an example when we prompt the text (an astronaut riding a horse on mars art-station, hd, dramatic lighting, detailed).

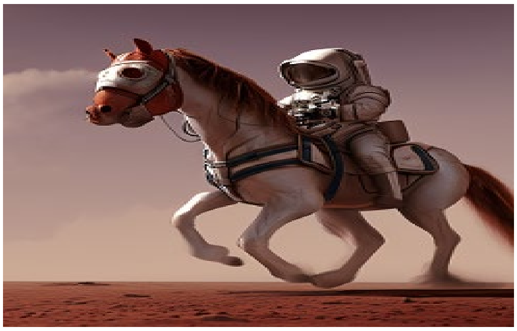


Fig.14 An image of an astronaut riding a horse on mars art-station, hd, dramatic lighting, detailed rendered with Stable diffusion model.

Here the model tested the power of prior by taking the base image and turning it into a project clip latent space and then they used PCA (Principal Component Analysis) to give them the highest points of interest. So, the important properties of the image have been learned correctly, and therefore we can use them in other applications without affecting other things. Of course, one of the uses of PCA is that it converts many properties into few properties, and this is what is applied in this model, so it is noticed that the most important properties were really related to the content of the basic image. Fig.15 shows examples when we prompt the text (kids with brown hair).

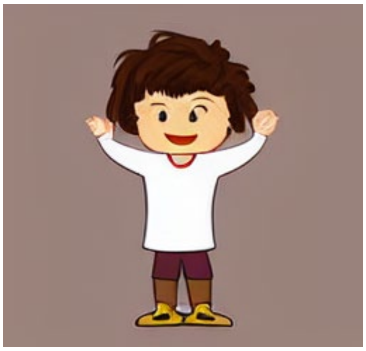


Fig.15 Character animation for kids with brown hair generated by diffusional Model.

Here they want to know how the decoder works. The first thing here they applied is that they took the sentence, and in our example, this is "a drawing of a green Pokémon with red eyes" in addition to the characteristics of the image, which is CLIP image embedding, which is linked to the given sentence, as we have seen how before. The result is shown in Fig.16.

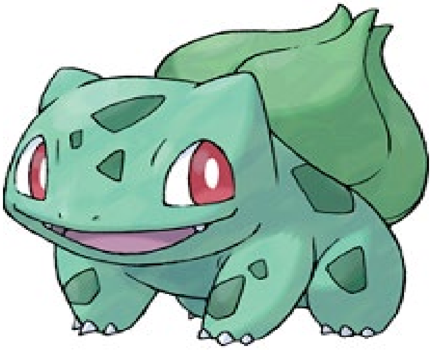


Fig.16 An image for drawing of a green Pokemon with red eyes generated by diffusional Model

Model 7: Stable diffusion model (SDM) v2

The Stable Diffusion Model 2.0 will be released by the end of Nov. 2022 and it bring big improvements and amazing new features that can be summarized as follows:

- New Text-to-Image Diffusion Models using a new OpenCLIP text encoder will bring better results that are closer to the prompt and are trained on an aesthetic subset of the LAION-5B. It can also have a base Resolution of 768 x 768 px.
- The new Super-resolution Upscale Diffusion will bring more Quality to upscales, but more GPU efficient and upscale to 2048 x 2048 px.
- The Models Depth-to-Image Diffusion Model will create depth maps that are important for coherent output, so you can change the prompt, but the composition of the image and the position of the elements in the image will stay the same.
- The In-painting Diffusion Model will bring improvements to in-painting, but also to out-painting, with the method.

5. Summary of Overall Comparison Between the Seven Systems

The seven text-to-image generation models examined in these papers were assessed using various datasets, including MS-COCO, Caltech-UCSD Birds, Oxford-102 Flowers, and others. These models employed different quantitative and qualitative metrics to evaluate their performance. Common quantitative metrics include Fréchet Inception Distance (FID), Inception Score (IS), and CLIP score, while qualitative assessments often involved human evaluations for photorealism and caption alignment. The FID score is the most universally applied metric, with lower scores indicating better quality and diversity in generated images. The Inception Score measures the clarity and diversity of generated images, and a higher score is better. CLIP score evaluates how well the generated image aligns with the provided text, with a higher score indicating better alignment.

For instance, LAFITE (Language-Free Image Text Embeddings) achieved an FID score of 8.12 on the MS-COCO dataset, highlighting its superior image generation quality. In contrast, models like DALL-E 2 and unCLIP also performed well, with FID scores around 10.39 and 10.87, respectively, showing their capability in generating diverse and realistic images. The models also varied in computational efficiency, with some requiring significantly more resources for training. Models like GLIDE and Make-A-Scene exhibited a balance between performance and computational efficiency, often achieving competitive FID scores with fewer parameters.

In my opinion, the best model overall would be GLIDE, as it achieves a strong balance between photorealism, caption alignment, and computational efficiency. While models like DALL-E 2 offer high performance, their computational demands make them less practical for broader use. GLIDE, with its relatively lower parameter count and high-quality outputs, presents a more efficient option without significant trade-offs in image quality or diversity.

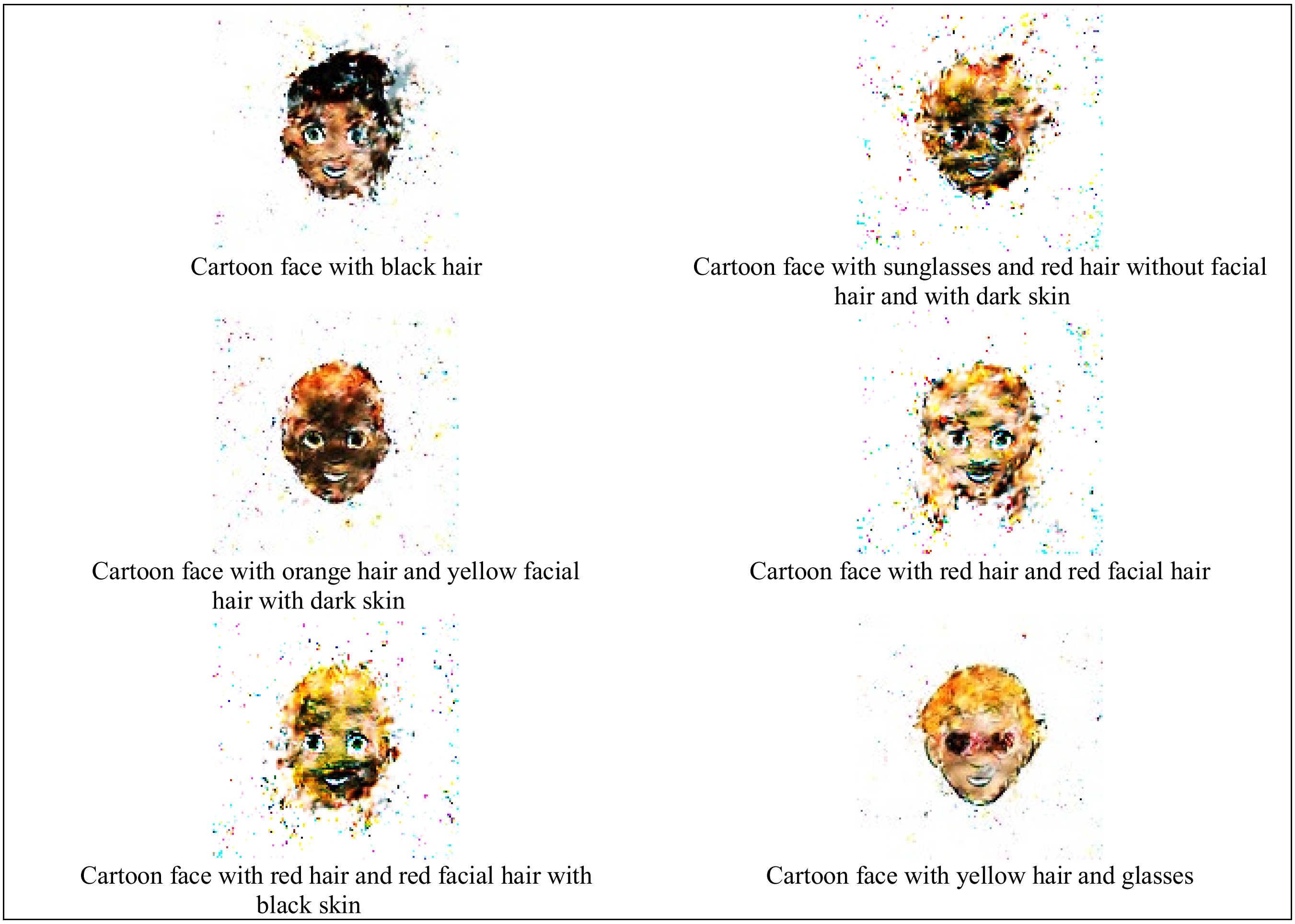
6. Comparison and result analysis

In this section, I will compare all the mentioned models by generating animation characters. under the same conditions and analyze the result using the same dataset that is the CartoonSet dataset.

6.1 Generative Adversarial Text to Image Synthesis

In this model the dataset[4] (Oxford-102 Flowers Dataset) doesn't contain many images for animation and cartoon characters so it can't be useful. Also, it concentrates on flowers and animals. For example: if I write (Cartoon face with black hair with different styles) then the model will generate the image as in Fig.17.

[4] Oxford-102 Flowers Dataset: This dataset contains 10000 high-resolution images of flowers belonging to 102 different categories. Each image is accompanied by a text caption describing the flower.



Cartoon face with black hair

Cartoon face with sunglasses and red hair without facial hair and with dark skin

Cartoon face with orange hair and yellow facial hair with dark skin

Cartoon face with red hair and red facial hair

Cartoon face with red hair and red facial hair with black skin

Cartoon face with yellow hair and glasses

Fig.17 A collection of character facial animations created by Generative Adversarial model.

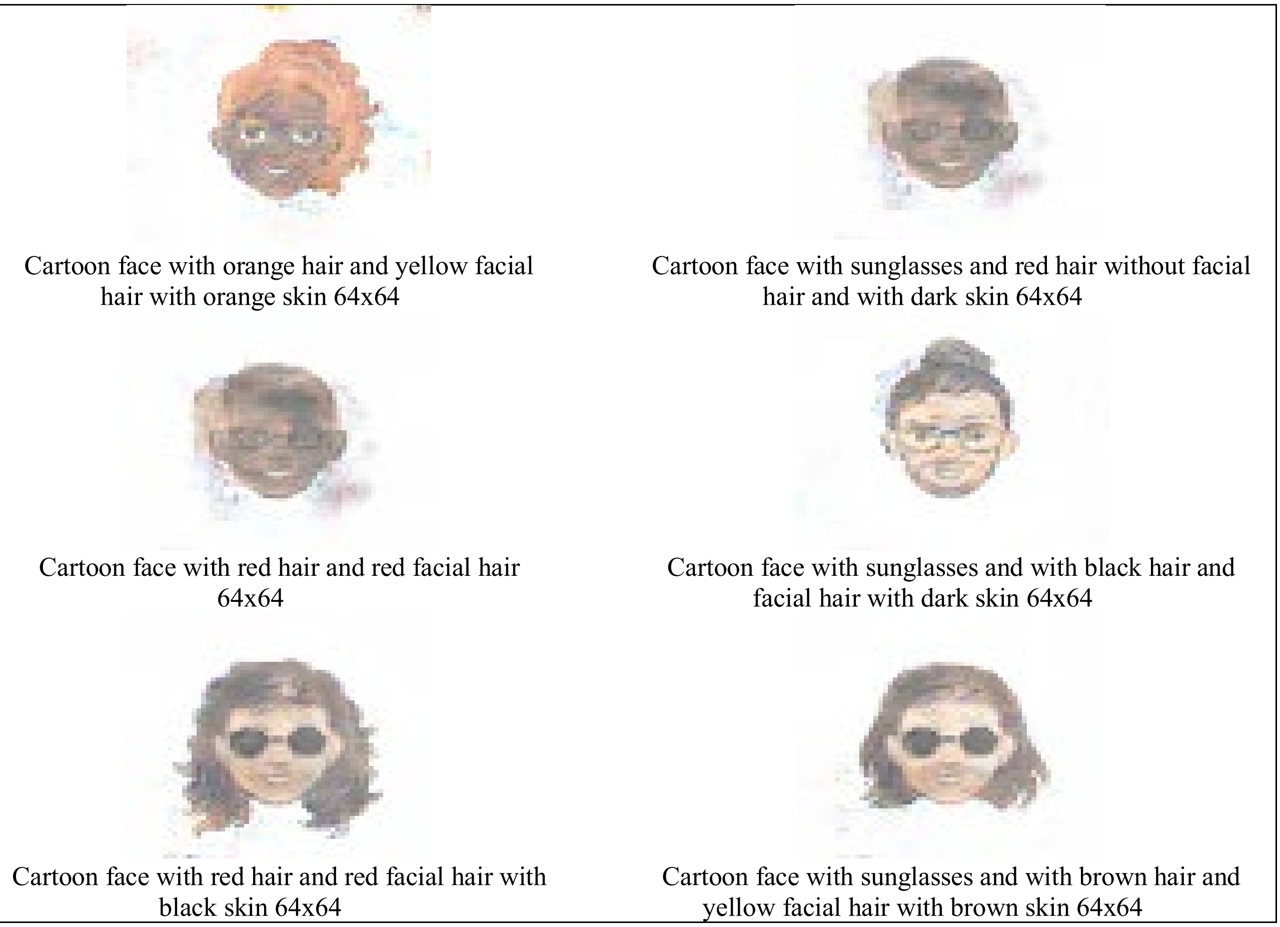
Here are the specific parameters for this example: The training duration is approximately 12 hours, the model runs for a total of 1000 epochs. The dataset consists of 10,000 images, each with a size of 128x128 pixels.

6.2 StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial[5]

This model dataset (that is Oxford-102 Flowers Dataset) lacks sufficient images depicting animation and cartoon characters, limiting its practical utility in that domain. Instead, its focus primarily centres around flowers and animals.

This model is a two-stage model where the First stage is 64x64 image and the output training succeeded in Fig.18 and the Second stage is Upsampling the image to 256x256 resulting in very bad output and “noise" images.

[5] https://arxiv.org/abs/1612.03242v2



Cartoon face with orange hair and yellow facial hair with orange skin 64x64

Cartoon face with sunglasses and red hair without facial hair and with dark skin 64x64

Cartoon face with red hair and red facial hair 64x64

Cartoon face with sunglasses and with black hair and facial hair with dark skin 64x64

Cartoon face with red hair and red facial hair with black skin 64x64

Cartoon face with sunglasses and with brown hair and yellow facial hair with brown skin 64x64

Fig.18 A collection of character facial animations created by StackGAN with 64X64 image size.

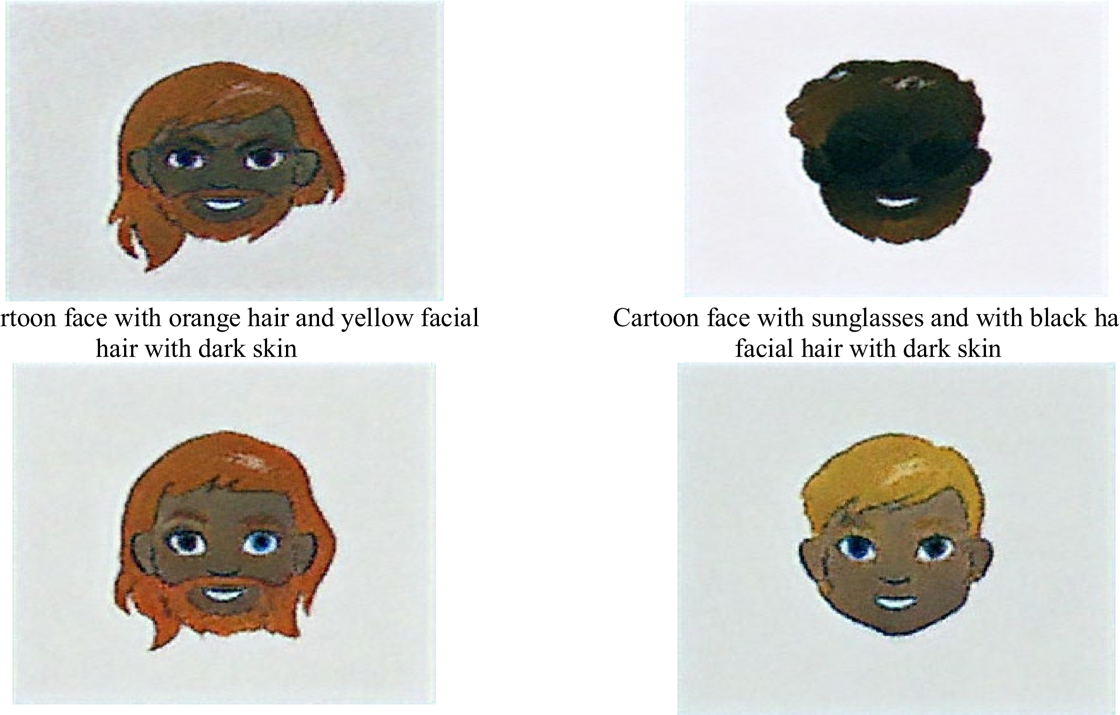
The parameters for this example are as follows: Training time is approximately 20 hours; the model is trained for a total of 1000 epochs. The dataset contains 10,000 images.

6.3 Hierarchical Text-Conditional Image Generation with CLIP Latent

In this model, the dataset contains many images for animation and cartoon characters so that it can be useful. Also, the dataset concentrates on character faces. For example: if I write (Cartoon face with black hair with different styles) then the model will generate some images as in Fig.19.

Cartoon face with black hair

Cartoon face with sunglasses and red hair without facial hair and with dark skin



Cartoon face with orange hair and yellow facial hair with dark skin

Cartoon face with sunglasses and with black hair and facial hair with dark skin

Cartoon face with orange hair and yellow facial hair with orange skin

Cartoon face without sunglasses and with yellow hair with yellow skin

Fig.19 A collection of character facial animations created by the CLIP Latent model

The parameters for this example are as follows: The training duration is approximately 17 hours, the model is trained for a total of 30 epochs. The dataset comprises 10,000 images, each with a size of 256x256 pixels.

6.4 GLIDE: Towards Photo-realistic Image Generation and Editing with Text-Guided Diffusion Models

This model's dataset lacks an adequate number of images depicting animation and cartoon characters, thereby limiting its practical usefulness in that particular domain. However, it primarily focuses on curating images of flowers and animals. For instance, if the prompt provided is "an oil painting of a corgi," the model is capable of generating an image as illustrated in Fig.20. This model has no implementation so we couldn't train the dataset.

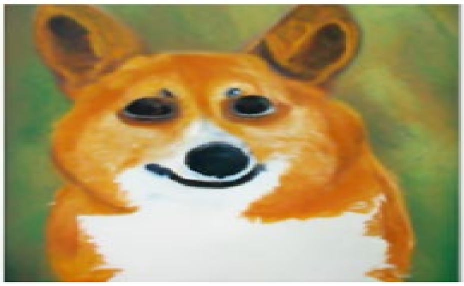
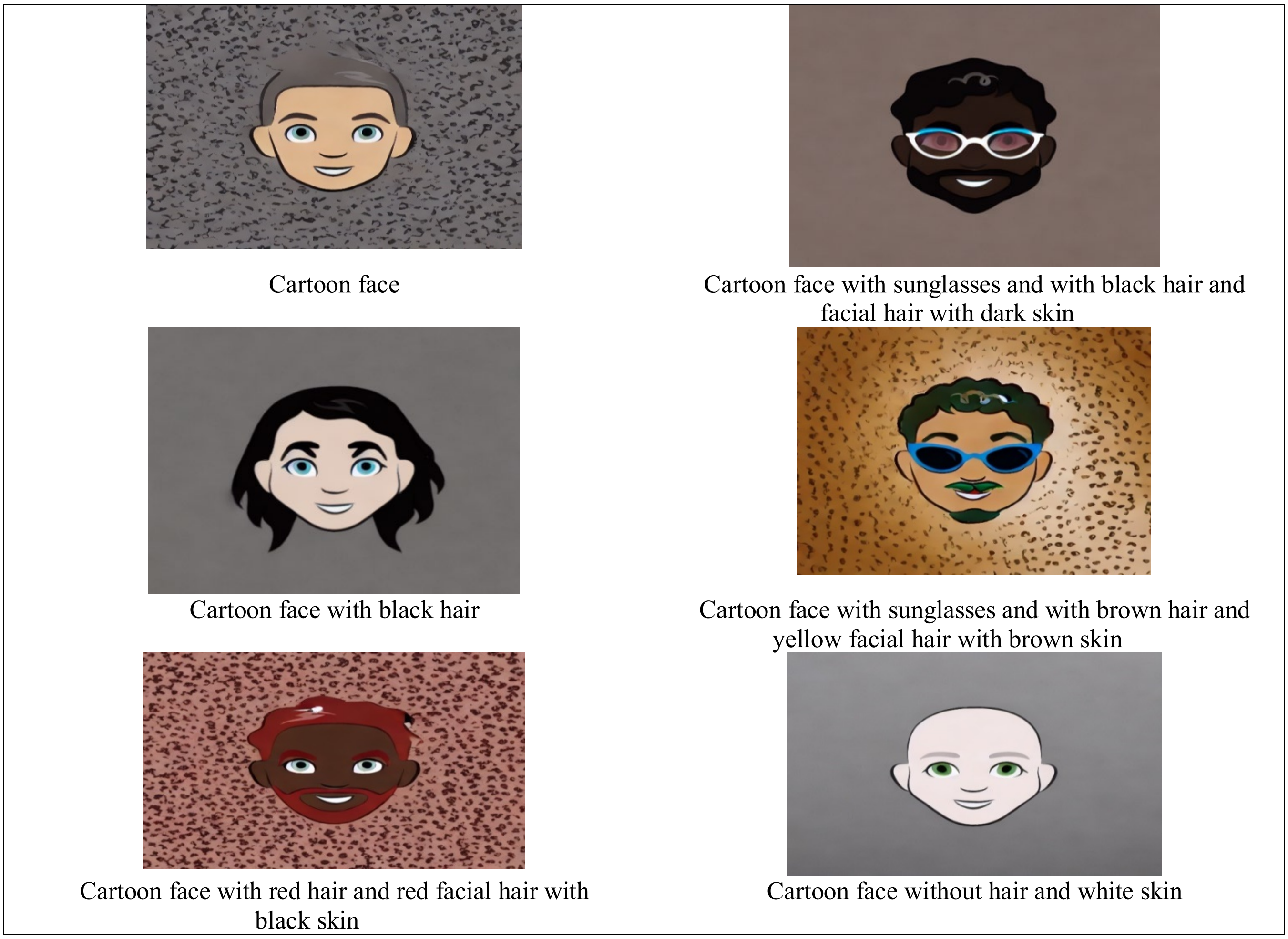


Fig.20 Animation Image generated by GLIDE

6.5 High-Resolution Image Synthesis with Latent Diffusion Models

This model benefits from a dataset abundant in images depicting animation and cartoon characters, making it highly useful in that regard. Additionally, the dataset primarily emphasizes faces and full-body characters. For instance, when provided with the prompt "(animation/cartoon kid with brown hair)," the model generates an image corresponding to the description, as depicted in Fig 21.



Cartoon face

Cartoon face with sunglasses and with black hair and facial hair with dark skin

Cartoon face with black hair

Cartoon face with sunglasses and with brown hair and yellow facial hair with brown skin

Cartoon face with red hair and red facial hair with black skin

Cartoon face without hair and white skin

Fig.21 A collection of character facial animations created by the Latent Diffusion Model.

The training duration is approximately 6 hours. The model is trained for a total of 6 epochs. The dataset consists of 10,000 images, each with a size of 512x512 pixels.

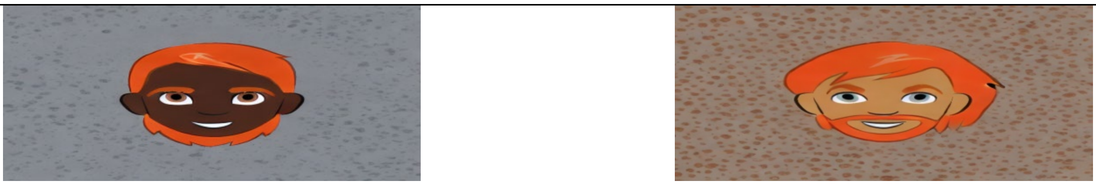
6.6 Stable Diffusion V1.0

Diffusion models for picture synthesis have recently taken the internet by storm. By the model in [37] with the title "High-Resolution Image Synthesis with Latent Diffusion Models", it has become evident that diffusion models are not only excellent at creating high quality, accurate images in response to a given prompt like Fig.22, but that this process is also significantly less computationally expensive than many competing frameworks.

However, there are many disadvantages and challenges ahead for Stable Diffusion that can be summarized as follows:

- Creating negative depictions of individuals or their environments, cultures, faiths, etc. that are humiliating, dehumanizing, or in any other way.

- Intentionally creating or spreading offensive stereotypes or discriminatory material.
- Posing as someone else without their permission.
- Sexual content that is posted without the viewers' permission.
- Errors and false information.
- Illustrations of shocking brutality and gore.
- Sharing of licensed or copyrighted content that violates the restrictions of that material's usage.
- Distributing content that violates the conditions of usage by changing copyrighted or licensed material.
- The LAION databases emphasize quantity over quality. The following clip retrieval images make it simple to validate the picture and text label pairings. You'll observe that the majority of the images are not accurately labeled, which will have a significant impact on the model's performance.



Cartoon face with orange hair and yellow facial hair with dark skin

Cartoon face with orange hair and yellow facial hair with orange skin

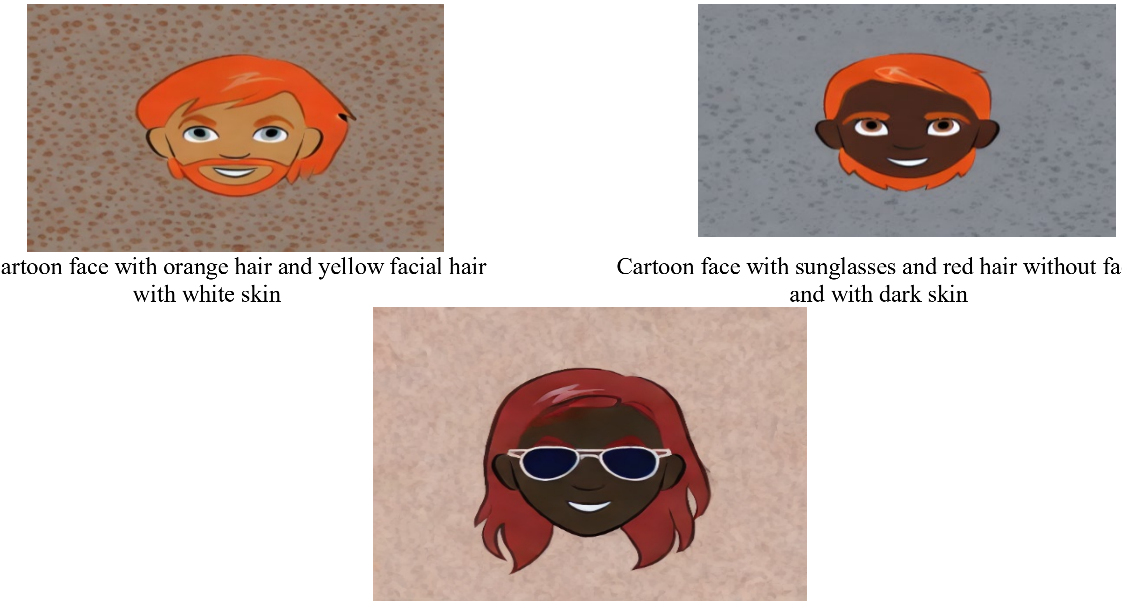
Fig.22 A collection of character facial animations created by Stable Diffusion V1.0.

The specifications for this example are as follows: The training time is approximately 6 hours. The model undergoes 6 epochs. The dataset includes 10,000 images, each with a size of 512x512 pixels.

6.7 Stable Diffusion V2.0

Stable Diffusion 2.0 represents a notable advancement over its predecessor, featuring substantial enhancements that have sparked mixed reactions within the AI community. While the upgraded architecture has been lauded for its innovative capabilities, there have been critiques regarding its application by Stability AI. A primary concern with SDM v2.0 is its demand for considerable processing time, making its deployment more time-intensive than earlier versions. Additionally, the preliminary outputs of SDM v2.0 might not fully match the quality of the final product, as they are prone to internal inconsistencies due to the layering technique utilized in the model.

Looking ahead, there's a strong possibility that OpenCLIP will develop a new text encoder based on the LAION-5B dataset, considering that Stable Diffusion is trained on subsets of LAION-5B. Given the pivotal role of the text encoder in the stable diffusion framework, any modification to this component could render obsolete much of the existing research related to prompt engineering. Furthermore, to accommodate the evolution of this technology, some of the current implementations will necessitate modifications to ensure backward compatibility between the older and newer versions. An illustration of the capabilities of this model can be seen in Fig.23, showcasing the kind of images it can generate.



Cartoon face with orange hair and yellow facial hair with white skin

Cartoon face with sunglasses and red hair without facial hair and with dark skin

Cartoon face with red hair and red facial hair

Fig.23 A collection of character facial animations created by Stable Diffusion V2.0.

For this example, the conditions are as follows: The training duration is approximately 6 hours. The model is trained for a total of 6 epochs. The dataset consists of 10,000 images, each with a size of 512x512 pixels.

7. Challenges

The domain of text-to-image generation technology, currently at the forefront of academic and professional interest, has been validated for its capability to instantiate contemporary design paradigms effectively and organically via textual directives. An extensive review of scholarly literature delineates the overarching influence of artificial intelligence, considering the objectives and breadth of existing methodologies, alongside its affirmative impact on stakeholders participating in multifarious architectural initiatives in Fig.39. This innovative technology's proficiency in decoding and rendering creative concepts from textual narratives is catalyzing a paradigm shift in the design and architectural sectors, enhancing efficiency and promoting inventive advancements.

This section delves into the complexities and limitations associated with prevalent evaluation methods and metrics in the context of generating animated characters from text descriptions.

- Viewpoint Diversity: A notable challenge in animating characters is accommodating multiple perspectives within a single film, as characters are often depicted from various angles. This complexity necessitates sophisticated modeling techniques to accurately render characters in a dynamic, multi-directional environment.
- Integration with Animation Software: A critical area for investigation is the compatibility between generated images and animation software, particularly concerning image format. Most animation tools are optimized for vector images, whereas many generative models produce raster images. Bridging this gap is essential for seamless application in animation workflows.
- Anatomical Considerations: The depiction of characters, especially the representation of hands with typically four fingers in cartoons, poses a unique challenge. This convention simplifies movement animation but requires careful attention in the model training process to ensure anatomical consistency with genre-specific expectations.

- Datasets: The development of datasets featuring dense cross-modal associations and captions grounded in visual content is crucial. Such datasets could significantly enhance the learning of compositional and fine-grained representations, thereby improving control over the image generation process tailored to animation character applications.
- Model Architecture: Future research should emphasize the exploration of text embeddings' quality and impact, the application of alternative generative models for animating characters from text, and the development of techniques that enhance scene understanding and context interpretation.
- Assessment Metrics: The refinement of evaluation metrics, including the Inception Score, Fréchet Inception Distance, R-precision, and Semantic Object Accuracy, is vital. These metrics have streamlined the assessment of models generating animated characters from text descriptions, but further enhancements are necessary to capture the nuanced success factors of such complex generative tasks [39-40].

8. Recommendations and Suggestions for Future Evaluation Metrics

Achieving precise alignment between an image and its corresponding text description poses a significant challenge, given the potential for ambiguity in the language used to describe visual content. Phrases like "semantically consistent," "fit," "match," or "properly represent" may all refer to the degree to which an image accurately reflects the provided text. An effective evaluation framework for text-based descriptions of animation characters should incorporate metrics that:

- Verify the presence and recognizability of mentioned objects within the image.
- Assess whether the objects in the input description are generated with accurate numerical and positional details.
- Confirm the precision with which the input description matches the generated image.
- Evaluate the model's sensitivity to minor modifications in the input description, such as word substitutions or the use of paraphrases.
- Provide explanations for any discrepancies between the input description and the generated image, enhancing the interpretability of the evaluation process [41].

In light of the current understanding of available metrics, we advocate for specific approaches to their application. Measuring the distance to the actual data distribution and assessing the visual quality of images via the Frechet Inception Distance (FID) is recommended. Furthermore, for evaluating the spatial accuracy of object placement within images, the use of FID for cropped objects is advised. It is crucial to detail the methodology behind score computation transparently, specifying whether baseline scores were recalculated or sourced from existing literature. Additionally, assessing the congruence between images and captions through Semantic Object Accuracy (SOA) and user studies is suggested to gauge image-text alignment comprehensively [42].

The rapid proliferation of generative Artificial Intelligence (AI) tools capable of producing lifelike images and text underscores the swift adoption of this technology. Platforms like DALL-E, Midjourney, and ChatGPT have captured widespread public interest, a phenomenon largely attributed to the extensive datasets compiled from varied online sources. These generative AI technologies are now contributing significantly to the expanding pool of data on the Internet, marking a notable evolution in the digital landscape [43].

9. Conclusion

The advent of text-to-image generation models, a groundbreaking facet of creative artificial intelligence, has introduced the ability to craft images from textual narratives. These models are progressively achieving results on par with the creations of professional artists and designers, sparking pivotal discussions on the future of creative roles, the potential for job displacement, and copyright issues, among other pertinent topics [44]. This research paper delves into the methodologies and techniques utilized in creating animation characters from textual descriptions. By conducting a thorough review of the literature and a comparative analysis of various methodologies, this study offers significant contributions to the domains of computer graphics and animation. The examination and comparison of existing methods for generating animated characters from textual descriptions have unveiled several challenges. One notable issue is the lack of detailed textual descriptions for images in databases, which is critical as most animated characters necessitate precise descriptions derived from written narratives. Additionally, the limitation of databases to only include character headshots rather than full-body descriptions poses a significant challenge, as full-body depictions are essential for animated film production. Another area of concern is the time and number of iterations required for character production, which highlights the need for improved efficiency in future algorithms. Moreover, the paper identifies the necessity for character outputs to be in vector format rather than pixelated forms, aligning with the preference for vector-based characters in animation software. Addressing these concerns will be crucial for advancing the field.

References

[1] Chitwan Saharia & William Chan, et al. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." arxiv Preprint arxiv:2205.11487 (2022).

[2] Fernandez, I. (2001) Macromedia Flash Animation and Cartooning: A creative guide. Berkeley, McGraw-Hill. PP. 47-93.

[3] X. Wu, K. Xu, P. Hall, A survey of image synthesis and editing with generative adversarial networks, Tsinghua Science and Technology 22 (6) (2017), pp. 660-674.

[4] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, Generative adversarial networks: An overview, IEEE Signal Processing Magazine 35 (1) (2018), pp. 53-65.

[5] Y. Hong, U. Hwang, J. Yoo, S. Yoon, how generative adversarial networks and their variants work: An overview, ACM Computing Surveys 52 (1) (2019) pp.1-43.

[6] N. Tiku, “AI can now create any image in seconds, bringing wonder and danger,” The Washington Post, Sep. 28, 2022. https://www.washingtonpost.com/technology/interactive/2022/artificial-intelligence-images-dall-e/ (accessed Dec. 10, 2022)

[7] L. Wang, W. Chen, W. Yang, F. Bi, F. R. Yu, A state-of-the-art review on image synthesis with generative adversarial networks, IEEE Access 8 (2020), pp. 63514-63537.

[8] A. Mogadala, M. Kalimuthu, D. Klakow, Trends in integration of vision and language research: A survey of tasks, datasets, and methods, arXiv:1907.09358 (2019).

[9] J. Agnese, J. Herrera, H. Tao, X. Zhu, A survey and taxonomy of adversarial neural networks for text-to-image synthesis, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2020).

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672-2680

[11] J. Agnese, J. Herrera, H. Tao, and X. Zhu, “A survey and taxonomy of adversarial neural networks for text-toimage synthesis,” Wiley Interdiscip Rev Data Min Knowl Discov, vol. 10, no. 4, p. e1345, Jul. 2020.

[12] T. Qiao, J. Zhang, D. Xu, and D. Tao, “Mirrorgan: Learning text-to-image generation by redescription,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2019- June, pp. 1505–1514, Jun. 2019.

[13] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv:1411.1784 (2014).

[14] Y. LeCun, C. Cortes, C. Burges, Mnist handwritten digit database (2010).

[15] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, in: International Conference on Machine Learning, 2016, pp. 2642-2651.

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

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111-3119.

[18] Z. S. Harris, Distributional structure, Word 10 (2{3}) (1954), pp. 146-162.

[19] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, M. Z. Afzal, Tac-gan - text conditioned auxiliary classifier generative adversarial network, arXiv:1703.06412 (2017).

[20] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, R. Urtasun, A. Torralba, S. Fidler, Skip-thought vectors, in: Advances in Neural Information Processing Systems, 2015, pp. 3294-3302

[21] H. Zhang, T. Xu, H. Li, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in Proceedings of the IEEE International Conference on Computer Vision, 2016, pp. 5907-5915.

[22] D. M. Souza, J. Wehrmann, D. D. Ruiz, Efficient neural architecture for text-to-image synthesis, arXiv:2004.11437 (2020).

[23] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: Fine-grained text to image generation with attentional generative adversarial networks, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2017, pp. 1316-1324.

[24] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Indian Conference on Computer Vision, Graphics \& Image Processing, 2008, pp. 722-729.

[25] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset, California Institute of Technology (2011).

[26] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, C. L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, 2014, pp. 740-755.

[27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next-generation image-text models. arXiv preprint arXiv:2210.08402, 2022.

[29] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, M. Z. Afzal, Tac-gan - text conditioned auxiliary classifier generative adversarial network, arXiv:1703.06412 (2017).

[30] Ayanthi, D.M. and Munasinghe, S. (2022) “Text-to-face generation with StyleGAN2,” Computer Science \& Technology Trends [Preprint].Available at: https://doi.org/10.5121/csit.2022.120805.

[31] Hao Wang et al. (2022) ManiCLIP: Multi-Attribute Face Manipulation from Text, arXiv.org. Available at: https://arxiv.org/abs/2210.0044v2(Accessed: Nov 28, 2022).

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

[33] Zhang, H. et al. (2017) “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” In: Proceedings of the IEEE international conference on computer vision. 2017. pp. 5907-5915.

[34] Zhang, H. et al. (2017) Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, arXiv.org. Available at: https://arxiv.org/abs/1612.03242 (Accessed: 08 June 2023).

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text conditional image generation with clip latent. arXiv preprint arXiv:2204.06125.

[36] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021

[37] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. “High-resolution image synthesis with Latent Diffusion Models In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. pp. 10684-10695.

[38] Chitwan Saharia & William Chan, et al. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." arxiv Preprint arxiv:2205.11487 (2022).

[39] Enjellina, E. V. Putri Beyan, and A. G. Cinintya Rossy, "Review of AI Image Generator: Influences, Challenges, and Future Prospects for Architectural Field," Journal of Artificial Intelligence in Architecture, vol. 2, no. 1, 2023.

[40] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: Fine-grained text to image generation with attentional generative adversarial networks, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1316-1324, 2018.

[41] T. Hinz, S. Heinrich, S. Wermter, Semantic object accuracy for generative text-to-image synthesis, IEEE transactions on pattern analysis and machine intelligence, Vol. 44(3), PP. 1552-1565, 2020.

[42] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, "GPTScore: Evaluate as You Desire", arXiv preprint arXiv:2302.04166, 2023.

[43] G. M. Ruiz de Arcaute, L. Watson, P. Reviriego, J. A. Hernández, M. Juarez, and R. Sarkar, "Combining Generative Artificial Intelligence (AI) and the Internet: Heading towards Evolution or Degradation?", arXiv preprint arXiv:2303.01255, 2023.
[44] V. Vimpari, A. Kultima, P. Hämäläinen, and C. Guckelsberger, "An Adapt-or-Die Type of Situation: Perception, Adoption, and Use of Text-To-Image-Generation AI by Game Industry Professionals," Proceedings of the ACM on Human-Computer Interaction,pp. 131-164, 2023.

المجل (١٢) العدد (١) (السنة ٢٠٢٥)

المجلة الدولية للحاسبات والمعلومات

متاح على الإنترنت على الرابط : /https://ijci.journals.ekb.eg

Egyptian Knowledge Bank
بنك المعرفة المصري

تحويل الوصف النصي إلى شخصيات رسوم متحركة: مراجعة ومقارنة لنماذج التوليد

[أ]سامح ظريف [ب] خالد امين [ج]عبد الفتاح النجار [د] مريان وجدى

[أ] قسم تكنولوجيا المعلومات،كلية الحاسبات و المعلومات- جامعة المنوفية، المنوفية، مصر

[ب] قسم تكنولوجيا المعلومات، كلية الحاسبات و المعلومات- جامعة المنوفية، المنوفية، مصر

[ج] قسم تكنولوجيا المعلومات، كلية الحاسبات و المعلومات- جامعة المنوفية، المنوفية، مصر

[د] قسم تكنولوجيا المعلومات، كلية الحاسبات و المعلومات- جامعة طنطا، طنطا، مصر

الملخص :

في الآونة الأخيرة، جذبت تقنيات الذكاء الاصطناعي التوليدية لتحويل النصوص إلى صور، والتي تُعرف عادةً بمولدات الصور بالذكاء الاصطناعي، اهتمامًا واسعًا بفضل قدرتها المذهلة على إنتاج الصور بسرعة استنادًا إلى الأوصاف النصية. وقد أثار هذا التطور وجهات نظر متباينة، خاصةً في قطاع الرسوم المتحركة، مما جعله محورًا للتحقيق العلمي بسبب إدخال الشبكات التوليدية الخصامية .(GANs) ورغم التقدم المحرز، فإن هذا المجال يواجه عدة تحديات تتطلب مزيدًا من البحث العلمي، مثل إنتاج صور عالية الدقة تحتوي على عناصر متعددة وإنشاء مقاييس تقييم تتماشى مع التقييم البشري. بالإضافة إلى ذلك، تكشف نتائج هذه الدراسة أن مولدات الصور بالذكاء الاصطناعي تمتلك القدرة على تعزيز الإنتاج الإبداعي بشكل كبير في مجال الرسوم المتحركة من خلال توفير مجموعة متنوعة من الشخصيات والبيئات بجودة بصرية فائقة. تهدف هذه المراجعة إلى فحص ومقارنة العمل المكثف في هذا المجال بشكل شامل، حيث ستوضح الخوارزميات المستخدمة، وتحدد المشكلات الحالية، وتقترح اتجاهات جديدة للبحث.