REVIEW ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2557

Towards Sustainable Image Synthesis: A Comprehensive Review of Text-to-Image Generation Models

Smita Bharne [a,*], Pallavi Sapkale [b], Ekta Sarda [b], Puja Padiya [a], Shamal Salunkhe [a]

[a] Department of Computer Engineering, Ramrao Adik Institute of Technology, D. Y Patil Deemed to be University, Navi Mumbai, India.

[b] Department of Computer Science and Engineering, Ramrao Adik Institute of Technology, D. Y Patil Deemed to be University, Navi Mumbai, India.

* Corresponding Author Email: smita146@gmail.com

DOI: https://doi.org/10.54392/irjmt2557

Received: 07-03-2025; Revised: 12-08-2025; Accepted: 07-09-2025; Published: 24-09-2025

Abstract: Text-to-image generation represents a rapidly evolving frontier in artificial intelligence, enabling the transformation of natural language descriptions into visually coherent and semantically rich images. This paper presents a comprehensive review of state-of-the-art generative models—including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and advanced Diffusion Models—focusing on their capabilities to produce high-fidelity, contextually accurate images from textual inputs. Additionally, we analyse leading sustainable image synthesis frameworks such as DALL-E 2, Stable Diffusion, Imagen, and MidJourney, assessing their advancements in image quality, semantic alignment, diversity, and computational efficiency. Our systematic evaluation highlights significant progress in generating realistic, high-resolution images while identifying persistent challenges related to semantic consistency, fine-grained control, ethical considerations, and substantial computational demands. We further discuss critical trade-offs between model performance and sustainability, fostering future research directions aimed at developing more efficient, fair, and environmentally responsible text-to-image generation systems. This survey serves as a guiding resource for the next generation of sustainable AI-driven text to image synthesis technologies.

Keywords: Deep learning, Diffusion model, DALL-E, Generative models, Text to Image generation.

1. Introduction

Text-to-image generation is in itself a change of paradigm for AI, combining linguistic understanding with visual creativity. It refers to the process of generating coherent and contextually appropriate images from textual descriptions, where the task puts to test the ability of models in capturing very fine-grained details of language and vision. Text-to-image generation is far from novel; it stands today at one of the crossroads of many vital subfields within AI: natural language processing, computer vision, and generative modelling [1]. Translating text inputs into visual outputs carries deep implications in so many areas. In the realm of creative work, text-to-image models are so radically new that they dramatically change the pace at which artists, designers, and, really, creators of all sorts can lay down concepts with a level of detail that is fast and until now unimaginable. A further field where these models become hugely important is in the creation of the environments—virtual and augmented reality—that are dynamically reacting to the user's input. They also hold promise in making things accessible; they might be able to create visual representations of textual information for the blind, making complex ideas easier to understand by interpreting and visualizing them [2]. Moreover, the task of text-to-image generation serves as a milestone on the horizon for AI improvement and tests the limits of the capability of current models to synthesize and integrate multimodal data.

Text to image generation has numerous applications beyond only image generation. AI generated images can be used for medical imaging purpose where real time images are difficult to find to identifying the rare diseases [3]. For fashion designing purpose a quick prototyping, personalized designs visualization in text to images system, speed up the product developments. The gaming industry can also use text-to-image systems to add the customized contents that augments immersive experiences [4]. The gaming industry employs text-to-image systems to produce dynamic assets, adding customized content that augments immersive experiences [5]. These models also used to assist to convert the written descriptions into visuals to enhance their communication ability [6]. In

recent years, deep generative models have developed quickly, including diffusion models, transformer-based structures, and generative adversarial networks (GANs). The fields like Digital art, virtual reality, education, have all been transformed by these models. [7, 8]. Thus this growing field also raise the question of sustainability in this domain. For sustainable text-to-image generation ethical issues, computing efficiency and environment impact are taken into consideration. These system must be developed responsibly. The definition of sustainability in this context is an integrated framework which makes the balance between the below two core pillars.

1. ***Computational Efficiency:*** Leading generative models often require large amounts of energy due to their heavy computing needs for training and operations. Most of the time these processes leave a significant carbon footprints [9]. There is a need to improve resources without losing image quality. By employing the training methods like model pruning, and efficient designs we can reduce the environmental costs and its implications [10, 11].
2. ***Ethical and Societal Considerations:*** For AI systems to be sustainable, fairness, accountability must be clear. Bias from the training dataset results into incorrect content generation [12]. This will maintain the ethical suitability and reducing the social damages. [13].

Although ethical consideration and environmental impact are important but computational efficiency is more important as will have direct impact on the scalability of text-to-image synthesis models. By considering this as a prime component, this paper offers a comprehensive review of the optimization methods and advancement support towards more sustainable generative AI.

1.1 Comparison with Prior Surveys and Contributions

Recent years have witnessed several surveys and reviews addressing aspects of text-to-image synthesis, focusing variously on model architectures, generative performance, or ethical considerations. However, these prior works typically exhibit limitations in scope and integration of sustainability dimensions. Current literature on text-to-image generation primarily concentrates on architectural innovations and perceptual quality improvements, often neglecting the complex trade-offs between model performance and sustainability considerations. Moreover, the absence of standardized sustainability metrics and frameworks impedes meaningful comparisons and holistic assessments. The Key distinctions and improvements of the present manuscript is highlighted in table 1.

This paper presents a comprehensive review of text-to-image synthesis models with distinctive focus on sustainability, an increasingly critical yet underexplored dimension in generative AI research. The major contributions and significance of this study are summarized as follows:

- Our survey contributes to this growing research area by providing comprehensive analysis of computationally efficient techniques for text-to-image synthesis, critically examining the trade-offs between model performance and sustainability.

Table 1. Comparison with existing literature survey

Key Aspects / Techniques	Coverage in Existing Surveys	Coverage in Our Survey	Representative Existing Literature (Citations)
Clear, Multidimensional Sustainability Definition (Efficiency + Environmental + Ethical)	✗	✓	[14,15]
Analysis of Ethical Issues (Bias, Fairness)	Partial	✓	[12]
Comparative Trade-off Analysis (Quality vs. Efficiency vs. Fairness)	✗	✓	[15]
Coverage of Diverse Model Architectures (GAN, Diffusion, Transformer, ImageGAN)	✓	✓	[16]
Integrated Future Research Roadmap (Multi-objective Optimization, Green AI, Ethical AI)	✗ / Partial	✓	[9](Partial), [17][✗]
Discussion of Efficiency Techniques (Pruning, Distillation, Sampling)	✗ / Partial	✓	[4](Partial),[18] [✗],
Explicit Focus on Sustainability Metrics in Model Evaluation	✗ / Partial	✓	[14] (Partial), [15], [✗]

- We critically analyse ethical concerns such as bias, fairness, and societal impact, integrating these discussions with computational sustainability to offer a unified evaluation framework.
- We outline future research directions and community-driven standards aimed at fostering sustainable development and deployment of text-to-image synthesis models, bridging current gaps in both practice and policy.

1.2 Significance of the study

This study is significant as it is interdisciplinary approach to understanding sustainability in text-to-image synthesis, an area composed for explosive growth and societal influence The review's comprehensive scope and actionable insights contribute toward establishing sustainability as a non-negotiable standard in future AI research and deployment.

1.3 Statistics of the AI generated Images

As per the statistics of the [19] more than 16 billion images are generated from the year 2022 to 2024. Figure 1 shows the statistics of the number of AI generated images till 2024. (* in figure indicates the number are increasing per year). After the launch of DALLE-2 an average of 34 million images are generated per day Adobe Photoshop's AI algorithms, Adobe Firefly, are the fastest-growing product. Three months after introduction, it created 1 billion photos. Midjourney has 15 million users, the most of any picture generating platform with publicly published information. As per Prodesigntools [20] reports 30 million users of Adobe Creative Cloud, which includes Photoshop, Adobe Firefly, and other graphic design and video editing software. 80% of the photos (12.590 billion) were made utilizing open-source Stable Diffusion models, services, platforms, and apps. This shows the popularity of the text to image generation models and platforms.

Presently with the rapid text-to-image creation breakthroughs have spawned numerous models, each claiming to push the limits. Image synthesis has a rich history, from GAN-based approaches to transformer- and diffusion-powered models. With so many models using diverse architectures, training regimes, and optimization strategies, it's hard to compare their merits and disadvantages.

2. Overview of Generative Architectures for Text-to-Image Synthesis

2.1 Evolution of Text-to-Image Generation Methods

The domain of text-to-image generation has seen big development: from simple and basic techniques to very complicated models that now generate highly realistic and contextually rich images. The following section traces the historical development of important breakthroughs in architectural innovation that shaped modern text-to-image generation [21, 22].

2.1.1 Early Approaches and Foundations

- Rule-Based Systems and Template Matching (Pre-2010): Preliminary endeavours in creating images from text were constrained to rule-based systems employing predetermined templates or rudimentary algorithms to correlate textual descriptions with static image components. These technologies were limited by their incapacity to produce original visuals or manage intricate descriptions.
- Statistical Models (2010s): During this decade, statistical models integrated more advanced language processing techniques; yet, they remained incapable of producing intricate and contextually precise visuals. The emphasis was largely on enhancing the system's textual comprehension, with minimal effect on visual quality.

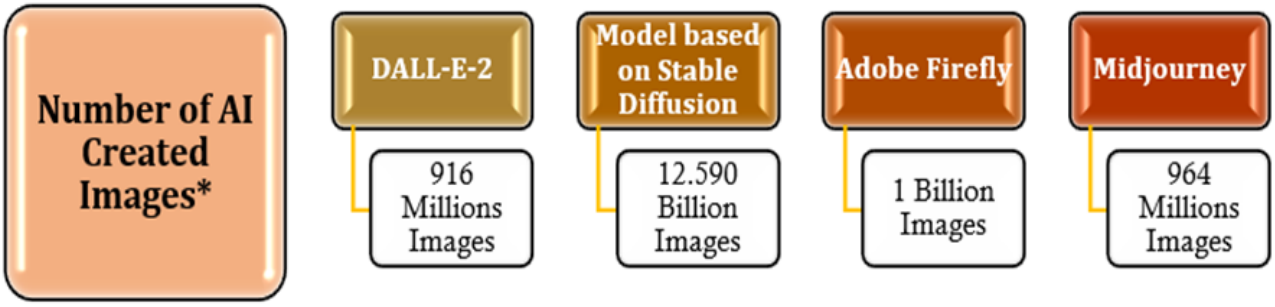

Figure 1. Statistics of the AI generated images

2.2 Evolution of Generative Models for Text-to-Image Synthesis

Recently, researchers have sought to develop entirely new content from scratch in areas such as text, images, and sound. This has been done by numerous strategies and tools developed, particularly deep learning techniques, which help in the implementation of generative models aimed at solving the difficult task of development of new material. The process of creating artificial material that is virtually indistinguishable from an underlying dataset is referred to as generative modelling. The potential application areas of generative models, which can deal with outcomes of multiple modes, are wide in the field of machine learning. This section delves into the four archetypical classes of generative models: Generative Adversarial Networks (GANs), Variational autoencoder (VAE) attention mechanism and transformers based architecture and diffusion models.

2.2.1 Generative Adversarial Networks (GANs):

The year 2014 saw an innovation by Goodfellow *et al.* in the form of Generative Adversarial Networks (GANs), which introduced a generator and discriminator attempting to outdo each other in producing high quality images. GAN mark a major advancement in image generation after introducing the adversarial training. It has potential improvements than the earlier methods. After this conditional GANs (cGANs) were introduced in 2016 which takes additional input like input descriptions from user. In 2017, Zhang et al. proposed StackGAN, a unique two stage model. In this model first stage produce a rough image which is refined in second stage in detailed with higher resolution. This multi-stage approach is a substantial progress in generating visually clear images from the text. Several studies have been carried out on GAN-based text-to-image synthesis since it was introduced in 2014, with remarkable progress being made in this field. Reed *et al.* [23] were the pioneers to research text-to-image synthesis through GAN based method from the ground established by Deep Convolutional GANs [24].

2.2.2 Variational autoencoder (VAE)

One way to look at it is that the variational autoencoder offers a framework for expressing an observation in a latent space as a probability distribution. The process begins with the input being encoded, where information is usually compressed into a smaller dimensional latent space. In general, autoencoders aims at minimizing the information loss during encoding and reconstruction process. Specifically, these should find a lower dimensional representation of an input matrix that's very larger leading to minimal reduction in content when reconstructed back again from its original form but compressed to lesser space [25].

2.2.3 Attention Mechanisms and Transformers

Transformers and Attention Mechanisms (2018): Attention methods greatly improved models' capacity to focus on diverse areas of input text. Xu *et al.'s* AttnGAN model aligned textual characteristics with visual aspects to improve picture quality and relevancy [26]. Transformers and Large-Scale Models (2019-2021): Transformer designs revolutionized text-to-image generation. Transformers, initially created for natural language processing, handled complex text-image interactions well. Top models like OpenAI's DALL-E used transformers to create detailed and artistic visuals from textual inputs, setting new benchmarks [27].

2.2.4 Diffusion Models and Photorealism

The creation of diffusion models was a big step forward in making high-quality, photorealistic pictures. Diffusion models, on the other hand, work by slowly improving a picture that is noisy into one that is clear and full of details. This method is used by Stable Diffusion and Imagen, two well-known programs, to make very realistic pictures while still letting you control the generation process [28]. DALL-E 2 and Beyond (2022): DALL-E 2 and similar models showed even better picture quality and understanding of context by using better transformer architectures and training methods to push the limits of what is possible in text-to-image generation.

Table 2 shows the comparative analysis of GANs, Attention Mechanisms, Transformers, and Diffusion Models in Text-to-Image synthesis.

From the literature it is observed that the there is a need of comprehensive evaluation across diverse Models. Limited subset of models was the main focus of several researches terminating on specific models in most situations. There is a call for a wide-ranging appraisal that takes in the whole gamut of contemporary models including DALL-E 2, Stable Diffusion, MidJourney, and Imagen for a total perspective of present-day capacities. In the next section we focus on the modern approaches to image synthesis.

2.3 Modern Sustainable Approaches to Image Synthesis

A number of state-of-the-art models have been created in text-to-image generation over the last couple of years greatly advancing research in this field. Complex architectures and techniques are what characterize the latest approaches at converting textual descriptions into various good-looking pictures that correspond to the reality. This section presents some of these best-known text-to-image models today.

Table 2. Comparative analysis of GANs, Attention Mechanisms, Transformers, and Diffusion Models in Text-to-Image Generation

Parameter	GAN (Generative Adversarial Networks)	Attention Mechanisms	Transformers	Diffusion Models	Citation
Core Architecture	Generator and Discriminator architecture, adversarial training	Focused on identifying important features within data through attention weights	Multi-layer self-attention and feed-forward neural networks	Probabilistic model that generates data by reversing a diffusion process	[1, 26, 29, 30, 31]
Text-to-Image Alignment	Indirect, needs explicit conditioning (e.g., via concatenation or auxiliary classifiers)	Improves alignment by focusing on key parts of the text relevant to the image generation	Strong alignment through transformer layers with direct language understanding	Diffusion process allows iterative refinement of image from noise based on text	[1, 26, 29, 30] 31]
Training Stability	Often suffers from instability, mode collapse	Typically, more stable when integrated with existing models (GANs, Transformers)	Relatively stable, but can require large datasets for training	More stable than GANs, but computationally expensive	[1, 26, 29, 30, 31]
Image Quality	Can produce high-quality images but prone to artifacts	Enhances image details by focusing on the most relevant elements in the text and image	Generates high-quality, often highly creative images with good coherence	High fidelity and photorealistic images with rich details	[1, 26, 29, 30, 31]
Scalability	Can be scaled but requires significant tuning	Highly scalable when integrated with large models	Scalable across various tasks, excels with large data	Scalable, but very computationally expensive, especially for high resolutions	[1, 26, 29, 30, 31]
Data Requirements	Moderate data requirements, but struggles with diversity in data	Enhances model performance without drastically increasing data requirements	Needs vast amounts of diverse data for optimal performance	Requires large and diverse datasets to learn diffusion and reverse processes	[1, 26, 29, 30] 31]
Computational Efficiency	Moderate, faster than transformers or diffusion models	Relatively efficient, adds a computational overhead but not prohibitive	Computationally intensive, especially for large models like GPT-based transformers	Computationally expensive due to the iterative refinement process	[1, 26, 29, 30, 31]
Customization	Allows for customization of outputs via conditioning	Provides fine-grained control over what aspects of the image to focus on	Provides strong customization by adjusting attention weights during generation	Less flexible but allows for gradual adjustments via diffusion refinement	[1, 26, 29, 30, 31]
User Accessibility	Accessible through many frameworks (e.g., TensorFlow, PyTorch)	Usually integrated into other models like GANs, Diffusion	Widely accessible, used in popular models like GPT, BERT	Increasingly accessible through frameworks like PyTorch and JAX	[1, 26, 29, 30, 31]

Ethical and Bias Concerns	Prone to biases in data, requires careful curation	Mitigates some biases by focusing on key textual cues but still requires attention	High risk of bias due to reliance on large-scale data without adequate filtering	Requires robust data filtering due to risk of overfitting biases	[1, 26, 29, 30, 31]
Strengths	Efficient for generating diverse image types, creative possibilities	Improves relevance of generated images, enhances text-to-image accuracy	Excellent at capturing the relationship between text and image	Produces high-quality, detailed images with strong alignment to text	[1, 26, 29, 30, 31]
Weaknesses	Stability issues, requires significant computational resources to avoid artifacts	Requires integration with other models, not standalone	Computationally expensive and complex to train at large scales	Highly resource-intensive, slow generation time due to multiple steps	[1, 26, 29, 30, 31]

2.3.1 DALL-E (OpenAI)

DALL-E is built on transformer models that produce images from textual cues by learning joint representations of text and images. It is a transformer based model which generate the images by combining the textual input with different modalities. It uses the Vector Quantized Variational Autoencoder (VQ-VAE-2) to convert the images in different tokens. These tokens are passed to transformer networks. This architecture produces the wider range of creative and sometimes unusual images beyond the conventional image generation approaches [32]. This model has a capacity to understand the complex prompt and generate the more creative images with high adaptability. Moreover DALL-E has strong semantic understanding which matches with exact input descriptions to make correct visual representations with deeper understanding of the input prompt.

2.3.2 DALL-E 2 -2022 (OpenAI)

DALL-E 2 has better quality and realism in pictures compare to earlier versions. The superior transformer architecture based on diffusion model is able to generate the images of good quality. Its photos have higher resolution than those generated by the previous version DALL-E and make them appear more like photographs due to realization for photo likeness of the originality. The new model equally includes abilities for image editing as well as inpainting allowing modification of particular regions within an image while obeying general appearance. DALL-E 2 displays greater image detail, better alignment with textual prompts, and can create clearer more authentic looking photographs [27].

2.3.3 Stable Diffusion (2022)

The Architecture of Stable Diffusion (SD) is a diffusion-based architecture that incrementally translates noisy images into clear, high-quality outputs. Unlike GANs, this technique has more reliable training dynamics and produces images that are closer to reality. SD has been described as being capable of producing high-quality images at high resolution, and has many uses across a large number of text descriptions that occur in natural language processing tasks. Its specialty is generating complex lifelike pictures in various artistic styles while accommodating differences in content category [33]. Due to several of its features such as production of more realistic images together with an open-source nature which also makes it valuable for researchers looking for tools which can help them in their work or anyone else interested in producing any type of visual art using tech solutions that exist today etc.; it's among the best choices for modelling human face appearance under normal working conditions. There are also utilities for refining and customizing models for specific tasks or aesthetics [28].

2.3.4 MidJourney

In 2022, MidJourney exploits a specific design that integrates transformer-based models and sophisticated generative techniques. The capability of generating imaginative and artistic images is supported by a massive dataset for training. MidJourney is particularly recognized for its artistic and imaginative abilities in creating beautiful visual works that defy conventional photo synthesis. It is adept at producing images using texts and instructions as stimuli. This approach to image generation with an emphasis on aesthetic and artistic aspect is unlike any other text-based recommenders for visualization apps. It is designed for creative professionals and artists who seek

visually appealing as well as morphologically varied images [34].

2.3.5 StackGAN

StackGAN is a Generative Adversarial Network for text-to-image synthesis. It produces images by reading text [30]. The unique thing about StackGAN is that it does this in two stages. The first stage involves producing general shapes and colours corresponding to the text while the second one refines this low-resolution depiction into a life-like high-resolution image that is enriched with details. Using dual phase architecture help StackGAN to increase quality of produced images as well as better to match them with the description given by making it quite good enough when detailed description through visual output is made from text. Table 3 provides the comparative analysis of the strength and weakness of each model discussed in this section.

3. Related Work

This part gives a summary of studies that provide information about models that generate images from texts. There are many types of models that generate images from texts and they have been extensively studied in the literature. Despite the wide coverage of text-to-image generative models, we summarized the related work of two main deep learning generative models: GANs and Diffusion Models in existing literature.

After being introduced in 2014, GAN-based text-to-image synthesis has been studied severally, something that has resulted in huge progress in this area. Reed et al. [23] were the first to look into text-to-image synthesis method, based on GAN, using convolutional neural networks as the basis for their work [24]. Stacked Generative Adversarial Networks (StackGAN) [30] proposed a two-step conditioning augmentation technique that enhances the variability of generated images and makes conditional-GAN training more stable. Based on the text description given, the first step GAN produces cheap low-resolution images that capture the basic shape and color characteristics of objects. The second step GAN generates high-quality realistic pictures (e.g., 256 x 256) based on outcomes provided by the first stage, but also by additional text information. MirrorGAN [35] introduces a structure from text-to-image-to-text with three models. In addition, it also recommends word sentence average embedding for global semantic coherence between the corresponding textual descriptions and generated images. An Image manipulator that matches images to texts that describe what is wanted from them, i.e. colour, feel and layout but untouching other things, is known as ManiGAN [36].

Table 3. Comparative analysis of the strength and weakness of each model

Model	Strengths	Weaknesses	Citation
DALL-E 2	- High creative potential, generating diverse and unique images. - Strong text-to-image alignment. - Produces high-quality, detailed images.	- High computational cost for inference. - Requires extensive fine-tuning for domain-specific tasks.	[32]
Stable Diffusion	- Efficient and scalable for various resolutions. - Open-source, customizable for different use cases. - Strong in photorealism and fine-tuning for style variations.	- Lower text-image alignment compared to DALL-E 2. - Requires technical expertise for optimization and tuning.	[33]
MidJourney	- Excels in creating artistic, stylized images. - User-friendly interface for non-technical users. - Great for creative exploration and visual artistry.	- Limited in producing highly realistic images. - Lacks flexibility for highly detailed, domain-specific tasks.	[34]
StackGAN	- Two-stage generation improves image quality. - Effective for generating higher resolution images (512x512 in Stage 2). - Strong text-to-image consistency in simple prompts.	- Struggles with complex text inputs. - Image diversity and quality are lower compared to modern models.	[30]

A combination of two crucial components constitutes ManiGAN. The first component integrates meaningful phrases into visual regions for better control. On the other hand, the second component deals with wrongly-matched properties as well as missing image parts. Despite the large number of researches conducted on this topic in English, a few cases are available in other languages. This paper introduces Attn-GAN (attention either way) mechanism which seeks to generate photo-realistic pictures from long descriptions written in Bangla text. In so doing, it synthesizes the most comprehensive features across all parts of an image, focusing particularly on relevant words or phrases given during such descriptions in human languages (natural languages) [37].

The study [38] describes an investigation on the application of transformer-based models (BERT, GPT-2, T5) for generating images from text, an unexplored domain in computer vision and natural language processing. Custom architectures are detailed which are intended for this purpose and constructed from them. Among the examined architectures of the transformer-based models on challenging datasets, it is found that T5 succeeds more effectively at making them seem visually beautiful as well as semantically meaningful [38]. Kang and colleagues developed a creative method of boosting GANs meant for making texts into visual images. They present in their paper a model called GigaGAN which shows how it can be done in a much better way when it comes to making those kinds of images. It is claimed that this new model produces images of high resolution and good quality quickly. When compared with other architectures developed before it, GigaGAN proves to be faster and capable of generating better looking pictures.

This is a huge leap forward in terms of utilizing GANs for enormous and complicated image synthesis jobs [39]. A solution was proposed for the widespread lack of combined text and shape data. CLIP-Forge needs just a pretrained image-text network such as CLIP and an unlabelled shape dataset when training is done in two steps. Utilizing a two-step training technique, one of the benefits of this method is that it does not rely on expensive inference-time optimization to generate various shapes for any text input etc [40].

Unlike how GAN approaches are done using small amounts of data, autoregressive models apply large amounts of data in generating text-to-image translations, including DALL-E by OpenAI [16] or Parti by Google [41]. Although they entail great computational complexity and time complexity because autoregressive models build on errors made in previous steps leading Sequential errors, a different case is diffusion models that have become popular in several different kinds of generating applications. The authors in [42] proposed CLIP-Forge to address the lack of matched data between texts and shapes, which is common in different datasets Establishing such a network is a two-stage process requiring just a pre-trained image-text model like CLIP as well as a no-label shaped database. One benefit this scheme has is that it can generate diverse shapes for a single sentence without the need for expensive heuristic search at runtime. In this study, CLIP-GEN, a self-supervised technique for automated text-to-image synthesis is introduced using the language-image priors obtained from a pre-trained CLIP model. In other words, you can instruct a text-to-image creator to operate using numerous pictures from a wide range of topics that are not annotated. By doing so, we will avoid having to amass large quantities of paired text-image material, which is too expensive to obtain. The method for generating images from text referred to as Imagen, is described in a paper published in [43]. Using just one encoder on sequential text and several dissipation models, it aims at creating high-definition pictures. In [44], Shi et al. introduced DiVAE, a VQ-VAE architecture with a diffusion decoder as the reconstructing component in image synthesis. They explored the possibility of embedding image data into the diffusion model to improve its performance and found that the latter can be achieved by slightly modifying U-Net.

4. Synthesis of Visual Performance, Computational Cost, and Ethical Considerations in Text-to-Image Models

While numerous text-to-image synthesis models have been proposed, their evaluation is often limited to visual fidelity neglecting broader sustainability and ethical dimensions. This section provides a critical comparative analysis of representative models—GAN-based, diffusion-based, and transformer-based—across multiple key criteria, offering a holistic perspective essential for sustainable AI development. Table 4 shows the analysis of the Text-to-Image Models on Key Dimensions.

4.1 Visual Quality and Fidelity

GAN-based models such as AttnGAN traditionally excel in producing sharp and coherent images but suffer from training instability and limited diversity due to mode collapse. Diffusion models like Imagen have recently set new standards in photorealism and sample diversity by progressively refining images through noise removal [15]. Transformer-based models leverage large-scale attention mechanisms to generate semantically consistent images that capture fine-grained text nuances, often outperforming GANs on zero-shot generalization tasks [45].

4.2 Computational Complexity and Efficiency

Training GANs generally demands less computational time relative to large transformers or

diffusion models, yet still requires significant GPU resources due to adversarial training cycles [9]. Diffusion models, while stable in training, entail longer inference times because of their iterative denoising steps, resulting in higher energy consumption [21]. Transformers demand the greatest computational resources both during training and inference, given their enormous parameter counts and self-attention mechanisms, contributing substantially to carbon footprints. Training large transformer models like DALL·E can consume megawatt-hours of electricity, raising sustainability concerns. Techniques such as pruning and knowledge distillation have shown promise in reducing the computational load of GANs by up to 40% without sacrificing quality illustrating pathways to greener AI [14, 17].

4.3 Ethical and Societal Considerations

Most models currently exhibit inadequate handling of bias and fairness. GANs and transformers trained on large web-scraped datasets risk reinforcing harmful stereotypes and producing biased content. Diffusion models, although relatively newer, have begun integrating bias mitigation strategies but lack standardized ethical evaluation protocols. The absence of comprehensive fairness auditing frameworks across architectures highlights a significant gap in sustainable model development. The ethical dimension is integral to sustainability but remains inadequately addressed in many text-to-image models [46]. Training data bias propagates into outputs, often reinforcing stereotypes related to gender, ethnicity, and culture [12]. Current mitigation efforts—such as dataset curation and rudimentary fairness constraints—are fragmented and lack systematic integration. The societal harms resulting from biased or misleading content underscore the urgency of embedding ethical frameworks directly into generative pipelines [47].

4.4 Trade-Offs and Sustainability in Text-to-Image Synthesis

There are intrinsic trade-offs between preserving ethical and sustainable computational practices and attaining high visual fidelity in text-to-image synthesis models. While GAN-based models have moderate energy requirements and comparatively faster inference, they have limited bias mitigation and training stability issues [9]. Diffusion models, like Imagen, use iterative refinement to produce better image quality and diversity, but their slow sampling results in high energy costs that affect environmental sustainability. Although transformer-based architectures offer zero-shot capabilities and strong semantic alignment, their large parameter counts result in significant energy consumption and carbon footprints [16]. Ethical issues are still not adequately addressed in any of these paradigms, indicating the need for integrated solutions that cooperatively maximize visual quality, resource efficiency, and equity.

Table 4. Critical Comparison of Text-to-Image Models on Key Dimensions

Model Type	Visual Quality	Training Cost (GPU-hours)	Inference Latency	Energy Consumption	Ethical Mitigation Efforts	Key Limitations	Citations
GAN (AttnGAN, DM-GAN variants)	High; capable of sharp images with some artifacts)	Moderate (~1200–1600 GPU-hours)	Fast; single pass inference	Moderate energy use	Limited bias mitigation, potential stereotyping	Training instability; mode collapse; insufficient ethics handling	[3, 7, 48, 49]
Diffusion (Imagen, Stable Diffusion)	Very High; state-of-the-art photorealism and diversity	High (~2500–3000 GPU-hours)	Slow due to iterative denoising	High energy footprint	Initial bias controls and fairness analyses emerging	Slow inference; large compute requirements	[14, 16, 49, 50, 51]
Transformer (DALL·E 3, Parti)	High; excellent semantic alignment and zero-shot performance	Very High (>3000 GPU-hours	Moderate latency; impacted by model size	Very High energy and carbon footprint	Limited but growing dataset filtering and fairness approaches	Large scale limits accessibility; ethical risks remain	[16, 17, 49, 50, 51]

When weighing the trade-offs between quality and cost, this comparative analysis shows that diffusion models outperform GANs in terms of visual fidelity but come with higher computational and environmental costs. Transformer-based models, although expressive, scale poorly in energy efficiency because of parameter explosion, which causes energy consumption to scale with model size. Few models incorporate thorough bias mitigation techniques with few ethical interventions; the majority rely on dataset-level filtering with little transparency.

5. Ethical Considerations in Text-to-Image Synthesis: Bias, Fairness, and Societal Impacts

This section follows the overview of model architectures and technical capabilities. It makes the point that sustainability is more than just being efficient and caring about the environment; it also means being morally responsible. Table 5 shows the analysis of the Bias and Societal Harms in Text-to-Image Generation.

5.1 Ethical Considerations in Text-to-Image Synthesis: Bias, Fairness, and Societal Impacts

Sustainability in AI systems exceeds beyond computational and environmental aspects to include significant ethical challenges. Text-to-image synthesis models, trained on extensive datasets sourced from the web, inherently embody the biases and stereotypes inherent in their training corpora [51]. These biases show up as distorted images of gender, ethnicity, age, and cultural identities, which could lead to harmful stereotypes and exclusionary stories in the images that are made [52]. Fairness in generative models is essential to ensure equal treatment of various social groups and without intentional countermeasures. Text-to-image systems may contribute to societal harms by generating misleading, offensive content that erodes user trust and deepens social divisions.

Recent research has started to address these ethical issues using techniques like dataset auditing, bias detection algorithms, and controlled generation methods [53]. However, these initiatives are still in their early phases, and comprehensive frameworks that incorporate moral principles into the creation and application of models continue to be an unexplored area of study. Given the societal impact and pervasive deployment of text-to-image technologies, embedding ethical considerations as core components of sustainable AI systems is imperative. This involves multi-stakeholder collaboration encompassing developers, users, policymakers, and affected communities to establish standards that safeguard fairness, accountability, and social good.

Table 5. Ethical Dimensions: Bias and Societal Harms in Text-to-Image Generation

Core Focus	Main Findings	Limitations	Citations
Gender and racial bias in generative models	Identified pervasive biases reflecting societal stereotypes in generated images.	Limited scope on text-to-image specifically; mostly language models.	[54]
Ethical risks of large language and multimodal models	Highlighted “stochastic parrots” problem and risks of amplifying harmful biases in multimodal datasets	Proposed general guidelines but lacked model-specific mitigation techniques	[17]
Dataset toxicity and harms in multimodal AI	Demonstrated that web-scraped datasets contain misogyny, hate, and bias that propagate into image generation	Suggested dataset curation but comprehensive bias mitigation remains open	[12]
Fairness interventions in vision-language models	Explored controlled generation to reduce stereotypical associations in text-to-image models	Methods limited in scalability and generalization across domains	[52]
Fairness and bias mitigation in diffusion models	Developed bias-aware training and dataset balancing for diffusion-based text-to-image synthesis	Early-stage research; lacks standardized benchmarks	[47]
Ethical challenges in large-scale transformers	Identified gaps in fairness and transparency in transformer-based image synthesis systems	Calls for integrating ethical frameworks throughout model lifecycle	[53]

6. Data Efficiency, Model Compression, and Lightweight Architectures for Sustainable Text-to-Image Generation

Sustainability in text-to-image synthesis is increasingly dependent on strategies that reduce data requirements, compress model size, and design inherently efficient architectures. These techniques collectively minimize computational costs and environmental impact while striving to maintain or improve image quality. Table 6 shows the different techniques and sustainability impact.

6.1 Data Efficiency

The volume and quality of training data profoundly influence both model performance and resource consumption. Recent advances in transfer learning, few-shot learning, and self-/semi-supervised pre-training have enabled models to learn effectively from limited labelled data, significantly reducing the computational burden of large-scale training [51]. Yang *et.al* [55] demonstrated that pre-training on unlabelled multimodal data followed by fine-tuning on smaller paired datasets achieves competitive synthesis quality while lowering training costs. Similarly, Kim *et.al* [5] applied few-shot learning techniques to accelerate adaptation to novel domains with minimal data, which is crucial for sustainable deployment in data-scarce contexts. Despite these benefits, data-efficient methods face challenges such as overfitting and reduced generalizability when training data is extremely limited or domain-shifted. Thus, balancing data efficiency with model robustness remains an open research area.

6.2 Model Compression Techniques

Model compression methods including pruning, quantization, and knowledge distillation have been extensively explored to shrink model size and accelerate inference. Pruning eliminates redundant weights or neurons, in GAN and diffusion models to reduce parameter counts by up to 40% without significant loss in image fidelity [56]. However, aggressive pruning risks degrading performance and requires careful tuning. Quantization reduces numerical precision of weights and activations, leading to memory savings and faster execution. Tao *et.al* [57] demonstrated quantization-aware training methods that preserve quality in generative models, albeit with dependency on hardware capabilities. Knowledge distillation trains smaller models to replicate the outputs of large models. After applied distillation to text-to-image transformers, achieving up to 60% model size reduction while retaining high semantic fidelity]. This technique, however, adds complexity due to the two-stage training process [58, 59]. These compression strategies significantly improve model efficiency, enabling deployment on resource-constrained devices and reducing energy consumption, essential for sustainability.

6.3 Lightweight Architectures

Designing lightweight generative architectures is an emerging frontier in sustainable image synthesis. Recent works explore efficient attention mechanisms and sparse transformers to reduce computational overhead in transformer-based models. MobileNet-style convolutions and efficient residual blocks that maintain representational power with fewer parameters in GANs and diffusion models. Architectural modifications like dynamic routing and adaptive sampling further optimize computation during inference. These innovations reduce floating-point operations (FLOPs) and model sizes substantially while sustaining or enhancing image quality, contributing directly to greener AI systems [60, 61].

Table 6. Summary of Techniques and Sustainability Impact

Technique	Description	Sustainability Impact	Challenges	Citations
Data Efficiency	Transfer/few-shot/self-supervised learning	Reduces labelled data and training time	Potential overfitting, domain shifts	[5, 51, 55]
Pruning	Removing redundant weights/neurons	Smaller models, faster inference	Possible accuracy loss	[56]
Quantization	Lower numerical precision weights	Memory reduction, accelerated inference	Hardware compatibility	[57-59]
Knowledge Distillation	Training compact student models	High compression, near-original quality	Additional training complexity	[58, 59]
Lightweight Architectures	Efficient attention, convolutions, blocks	Lower FLOPs, energy consumption	Requires novel design and tuning	[60, 61]

Integrating these techniques within text-to-image pipelines presents promising paths to reconcile high visual quality with sustainability. However, current approaches often address these methods in isolation. Additionally, developing standardized benchmarks to evaluate trade-offs between model size, speed, energy consumption, and image quality will be crucial for progress.

7. Environmental Implications of Computational Resources in Text-to-Image Generation

The need for computational resources has significantly increased in combination with the recent increase in the capability and use of text-to-image synthesis models. This increase has a direct impact on the environment, especially when it comes to energy use and related carbon emissions. For the responsible development and application of generative AI technologies, it is essential to understand and evaluate these implications.

7.1 Computational Resource Demands, Environmental Impact and Energy Footprint of Training and Inference

Due to the rapid advancements in the text- to-image synthesis models requires the high amount of computational energy which leads to the negative impact on environment. By reducing the effects of these can enhance the sustainability of the generative AI. Assessment of the text to image models also involves their computational and environmental cost. AttenGAN offering the faster inference to the prompts with reasonable training costs. Diffusion models like Imagen produces the highly accurate image but requires the more computational cost and power due to iterative sampling. Transformer based models like DALL E-3 suffers from more energy consumption and carbon emission due to their large size and complexity, although it has superior semantic fidelity. By using specialized hardware structure these models trained on larger datasets over the period of time. To train the DALL-E model the energy consumption is over the 6 MWh of electricity which results into the approximately emission of 2,500 kg of carbon emission [16, 17]. The diffusion models also consume the 2 to 5 Mwh of energy due to their complex denoising methods [14, 15].

GAN-based models generally have lower energy requirements, typically less than 1.5 MWh, but still contribute noticeably to the overall environmental burden [39]. While training energy dominates initial carbon footprint assessments, inference—the generation of images at deployment scale—accumulates considerable energy use, especially for diffusion models that require multiple sampling steps. Large-scale applications involving real-time or batch image generation amplify this footprint, often surpassing training energy over a model's operational lifetime. This aspect is crucial for sustainability but remains insufficiently addressed in many evaluations. Table 7 summarizes the estimated energy consumption and carbon emissions associated with representative models, based on recent literature and energy emission benchmarks.

- **Energy Consumption** is often measured in megawatt-hours (MWh) during training.
- **Carbon Emissions** (CO_2 equivalents) depend on the energy source; estimates assume average grid emission factors.

7.2 Implications for Sustainable AI Development

Quantifying environmental costs informs trade-offs essential for sustainable AI. While diffusion and transformer models advance image quality and semantic understanding, their energy demands challenge ecological responsibility. GANs offer a more resource-efficient alternative but with compromised stability and output diversity. This necessitates integrated strategies including model compression, data-efficient learning, and architecture optimization (discussed in Section 3.X on model efficiency), as well as advocacy for renewable energy adoption in data centres. Techniques such as model pruning, quantization, knowledge distillation, and hardware accelerators have been developed to reduce both training and inference energy consumption without substantial quality loss. Moreover, the adoption of renewable energy sources in data centers can further alleviate the carbon footprint of model training and deployment. Complemented by prior literature, emphasizes the need for the community to adopt standardized energy and carbon metrics. Integrating environmental considerations into model design and evaluation frameworks is imperative to ensure the responsible and sustainable growth of text-to-image generation technologies.

8. Evaluation Metrics for the Image Generation Models

Many metrics used to rate the quality of a model today majorly evaluate two things; how good the images are and the extent to which text is relatable with images. Frechet Inception Distance (FID) [61] and Inception Score (IS) [62] are examples of such metrics that are used widely for assessing model image quality. For a more detailed explanation, additionally, we use Clip Score method [62] and resolution to evaluate common sense.

Table 7. Estimated Energy Consumption and Carbon Emissions of Representative Text-to-Image Models

Model Type	Approximate Training Energy (MWh)	Estimated Carbon Emissions (kg CO_2 eq.)	Findings	Sources
GAN (AttnGAN)	~0.5 – 1.5	~250 – 750	Moderate compute, lower training time	[9]
Diffusion (Imagen)	~2 – 5	~1000 – 2500	High compute due to iterative denoising	[14, 16]
Transformer (DALL·E 3)	>5	>2500	Large-scale transformer models with billions of parameters	[16, 17]

Table 8. Quantitative comparison of modern text-to-image generation models based on evaluation metrics

Model	Inception Score (IS)	Frechet Inception Distance (FID)	CLIP Score	Resolution (px)	Remarks	Citation
DALL-E 2	4.85	10.39	0.94	1024x1024	Excels in creativity and contextually aligned image generation.	[32]
Stable Diffusion	4.56	12.63	0.93	512x512 (default), scalable	Strong in photorealism, customizable via fine-tuning.	[33]
MidJourney	4.32	13.87	0.92	1024x1024	Known for its artistic and stylized image generation.	[34]
Imagen	5.16	7.27	0.95	1024x1024	Best in photorealism, high fidelity and detailed image synthesis.	[63]
StackGAN	3.70	74.05	0.84	256x256 (stage 1), 512x512 (stage 2)	Two-stage generation improves image quality, but resolution is lower.	[30]

a) *Inception Score (IS)* has been designed specifically for assessing the quality and variety in images produced when using this concept. Measures the diversity and quality of generated images. Higher is better. Basically, this concept looks at measuring the “certainty” of a classifier while it makes predictions of labels for generated images and their spread or diversity across all images’ dataset.

b) *The Frechet Inception Distance (FID)* is yet another way to assess the standard of artificial images made by determining how close they resemble real images in terms of their distribution. Tested against a different evaluation method called Inception Score (IS) that only evaluates quality and diversity of single image, FID measures specific statistical features within these samples against what would be expected from any sample in reality

c) *The CLIP score* quantifies how similar the generated image is an initial text input. CLIP is a program designed to comprehend the rapport between words and pictures. It encodes them all in a shared embedding field; hence, similar text should appear near pictures that share similar meanings.

d) *The resolution* of the images that are produced refers to their dimensions. This is usually defined in pixels and is expressed as width px by height px. The resolution has a significant impact on how good an image looks to the human eye and the number of details that they contain.

Table 8 gives the Here is a quantitative comparison of modern text-to-image generation models based on commonly used metrics such as Inception Score (IS), Fréchet Inception Distance (FID), CLIP Score, and Resolution.

9. Comparative Analysis of the Methodological Approaches in Text-to-Image Generation

The process of generating images from text involves converting natural language descriptions into visually meaningful images using sophisticated machine learning methods. Figure 2 illustrates the architecture of the text-to-image generation model. It starts with a Text Encoder, which transforms the input text into numerical embeddings that represent its semantic meaning. These embeddings are then projected into a latent space via a neural network, aligning them with the visual domain. From there, an Image Generator based on different models like GANs, diffusion models, or VAE creates an image based on the latent representation. The result is a generated image that is both contextually accurate and visually detailed, closely matching the input text.

Algorithm

Inputs:

- Training image x_0 and corresponding text prompt y
- Number of diffusion steps T
- Noise schedule $\{\beta_t\}_{t=1}^{T}$
- Model parameters θ (for ϵ_θ)

Precomputations:
For $t = 1, \ldots, T$

1. $\alpha_t = 1 - \beta_t$
2. $\alpha_t = \prod_{s=1}^{t} \alpha s$

Training Phase:

For each training iteration**:**

1. Sample an image x_0 and its text prompt y from the training set.
2. Sample a diffusion time step $t \sim$ Uniform$\{1, \ldots, T\}$
3. Sample noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
4. Compute the noisy image:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon.$$

5. Predict the noise using the model: $\epsilon_\theta(x_t, t, y)$
6. Compute the loss:

$$L(\theta) = \mathbb{E}_{t, x_0, \epsilon}[\| \epsilon - \epsilon_\theta(x_t, t, y) \|^2].$$

7. Update the model parameters θ via backpropagation to minimize $L(\theta)$

Inference Phase:

1. Given a text prompt, y encode it appropriately.
2. Initialize $x_T \sim \mathcal{N}(0, \mathbf{I})$ (a sample of pure noise)
3. For $t = T, T-1, \ldots, 1$
4. Use the trained model to predict the noise. $\epsilon_\theta(x_t, t, y)$
5. Compute the parameters (mean $\mu_\theta(x_t, t, y)$ and variance $\Sigma_\theta(x_t, t, y)$ of the reverse distribution:
6. $p_\theta(x_{t-1} \mid x_t, y) = \mathcal{N}(x_{t-1};\, \mu_\theta(x_t, t, y), \Sigma_\theta(x_t, t, y))$.
7. Sample x_{t-1} from $p_\theta(x_{t-1} \mid x_t, y)$
8. Output x_0 as the generated image conditioned on the text prompt y

9.1 Dataset Details

For comparative analysis we have used the input text description from the four different dataset CUB-200, FashionGen, LAION-5B and DiffusionDB, [64-67]. All datasets are large-scale, publicly available dataset created to advance research and development in generative AI. It focuses on text-to-image generation using diffusion models. The dataset provides a diverse collection of text prompts paired with their generated images, offering a valuable resource for evaluating and fine-tuning diffusion-based generative models. It has over 14 million plus text-image pairs. The average prompt length is 10-20 words.

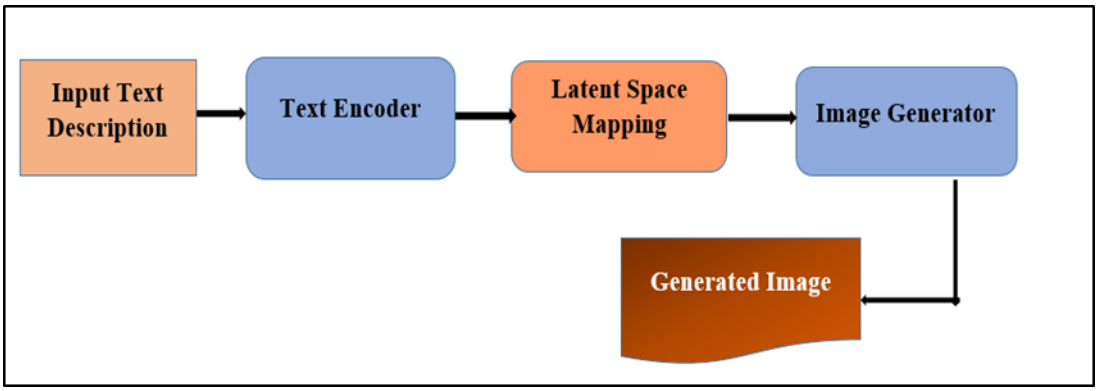

Figure 2. Text to image generation model

Table 9. Dataset details

Dataset	Domain	Number of Samples	Average Prompt Length (words)	Image Resolution	Description
CUB-200	Fine-grained bird species	~11,788	10–15	224×224	Rich textual descriptions per image
FashionGen	Fashion products	~293,000	12–18	256×256	Multi-modal annotations
LAION-5B	General web images	~5 billion pairs	Varies (avg. 10–20)	Variable	Large-scale, diverse web data
DiffusionDB	Generated image prompts	~2.8 million	15–25	512×512	Curated prompt gallery for diffusion models

The pre-processing text is tokenized using a pre-trained language model tokenizer BERT. After that the images are resized to a uniform dimension, 512×512 for model compatibility. After that we compare the results by the following models DALL-E, Stable Diffusion, Imagen and MidJourney. Table 9 give the details of dataset.

The Frechet Inception Distance (FID) and The CLIP score is used to evaluate the performance of the model.

Mathematical formula for FID (Frechet Inception Distance) is given by equation (1)

$$FID = || \mu_r - \mu_g ||^2 + Tr\left(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}\right) \quad (1)$$

where μ_r, Σ_r are the mean and covariance of the real data features, and μ_g, Σ_g are those of generated data.

CLIP Similarity Score: Measures alignment between generated images and text embedding. Mathematical formula is given below.

$$S_{CLIP}(x, y) = \text{cosine\_similarity}\left(\text{CLIP}(x), \text{CLIP}(y)\right) \quad (2)$$

9.2 Results Analysis and Discussions

The performance of each model is evaluated based on the FID and CLIP Score. Figure 3 shows the comparative graph analysis of the FID score vs CLIP score for each dataset. The measurable indications of FID score and CLIP score described below.

- ***FID Score (↓):*** Lower is better, indicating closer alignment to the real image distribution.
- ***CLIP Score (↑):*** Higher is better, indicating better semantic alignment between text and generated images.
- ***Diversity score:*** is a metric measuring variance in generated images, used to evaluate if a model produces varied outputs instead of similar ones for different or identical prompts.

Figure 3 shows the graphical representation of the FID vs CLIP score per Dataset. Across all the sets, Imagen consistently shows the optimal performance with the lowest FID, highest CLIP scores, and high diversity, indicating better image quality and alignment with the text. Stable Diffusion and MidJourney show competitive but slightly poorer performances, while DALL-E consistently lags behind the rest. The marker size corresponds to the diversity score (larger markers indicate higher diversity). High diversity is desirable as it reflects a model's ability to generate a wide range of distinct and creative images from similar or identical prompts, enhancing user experience and practical applicability. Models like Imagen and Stable Diffusion that couple high diversity with strong FID and CLIP metrics present a favourable trade-off between quality, semantic consistency, and creativity. Lower diversity, as observed in DALL·E, may limit applicability in contexts requiring varied or novel image generation, despite acceptable semantic alignment. Incorporating diversity metrics alongside traditional FID and CLIP scores provides a more holistic evaluation of text-to-image models. This understanding is essential for advancing models that meet both quantitative benchmarks and qualitative creative expectations. Table 10 shows the tabular comparison of the each model performance on FID, CLIP and diversity score values across all the datasets along with brief performance analysis.

Figure 4 shows the FID and CLIP score per dataset for each model. Amongst all the datasets Imagen has lowest FID score of 23 for LAION-5B dataset reflecting high image quality and DALL-E have the highest FID score of 50 for FashionGen dataset. From figure 4 we can analyses that for all the datasets Imagen has highest CLIP score of 0.36 LAION-5B dataset indicating superior for generated images while DALL-E tends to show lower performance for FashionGen dataset.

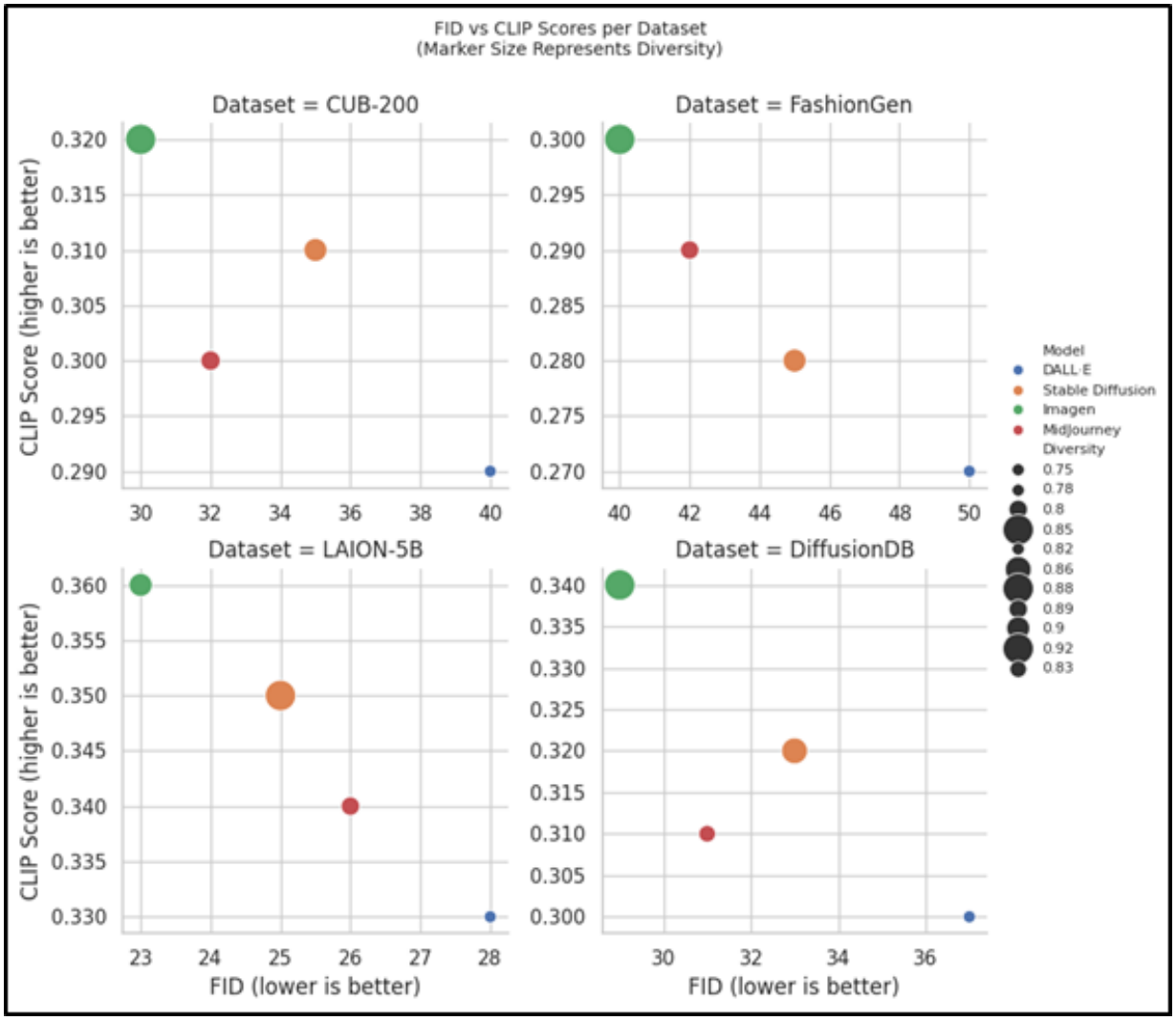

Figure 3. FID vs CLIP score per Dataset

Table 10. Performance analysis of each model on FID, CLIP and diversity score

Dataset	Model	FID	CLIP	Diversity	Analysis
CUB-200	DALL-E	40	0.29	0.75	Highest FID and lowest CLIP; underperforms in both quality and alignment.
	Stable Diffusion	35	0.31	0.80	Better than DALL·E; moderate performance with improved quality and alignment.
	Imagen	30	0.32	0.85	Best performance: lowest FID, highest CLIP, and strong diversity.
	MidJourney	32	0.30	0.78	Moderate performance; competitive but slightly behind Imagen.
FashionGen	DALL-E	50	0.27	0.78	Highest FID and lowest CLIP; indicates the poorest performance on this dataset.
	Stable Diffusion	45	0.28	0.82	Moderate performance; better than DALL·E yet still lagging behind the top performer.
	Imagen	40	0.30	0.86	Best performance with the lowest FID and highest CLIP, along with strong diversity.
	MidJourney	42	0.29	0.80	Shows moderate performance; comparable to Stable Diffusion but slightly inferior to Imagen.
LAION-5B	DALL-E	28	0.33	0.88	Competitive values; however, still not reaching the optimal zone seen in top performers.

	Stable Diffusion	25	0.35	0.92	Very good performance with low FID and high CLIP and diversity scores.
	Imagen	23	0.36	0.90	Best performance overall: optimal FID, highest CLIP, and strong diversity.
	MidJourney	26	0.34	0.89	Slightly behind Imagen; overall competitive but with a small gap in quality/alignment.
DiffusionDB	DALL-E	37	0.30	0.82	Highest FID and lowest CLIP in this dataset; underperforms relative to others.
	Stable Diffusion	33	0.32	0.86	Moderate performance; better than DALL·E but not the best.
	Imagen	29	0.34	0.88	Best performance with the lowest FID and highest CLIP and diversity; leads the pack here.
	MidJourney	31	0.31	0.83	Moderate performance; slightly trailing Imagen in all key metrics.

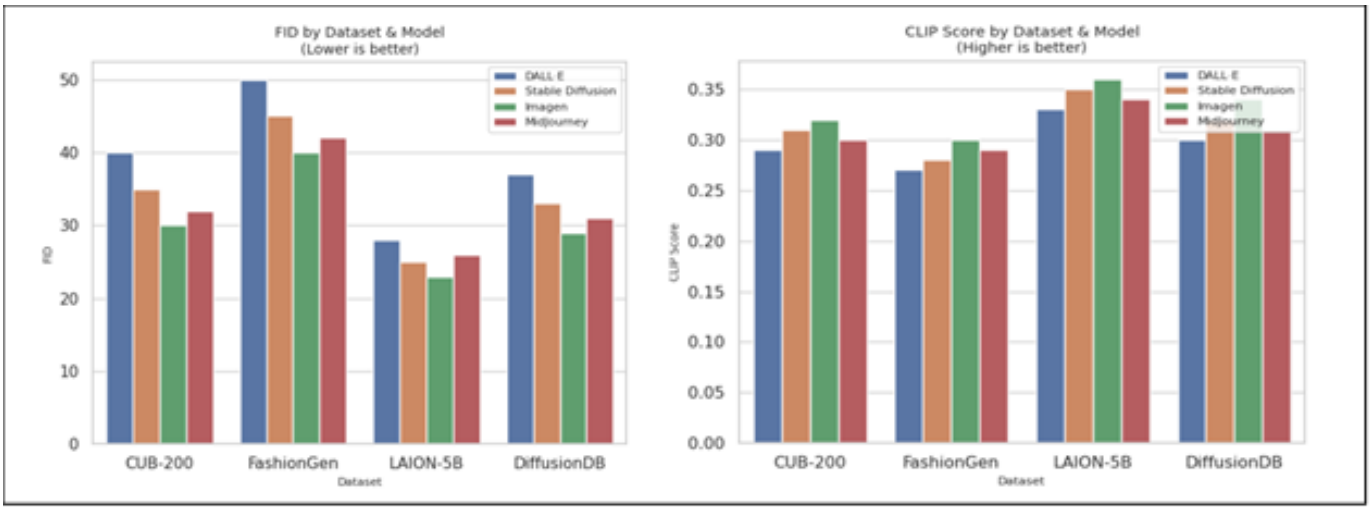

Figure 4. FID and CLIP score for each model and dataset

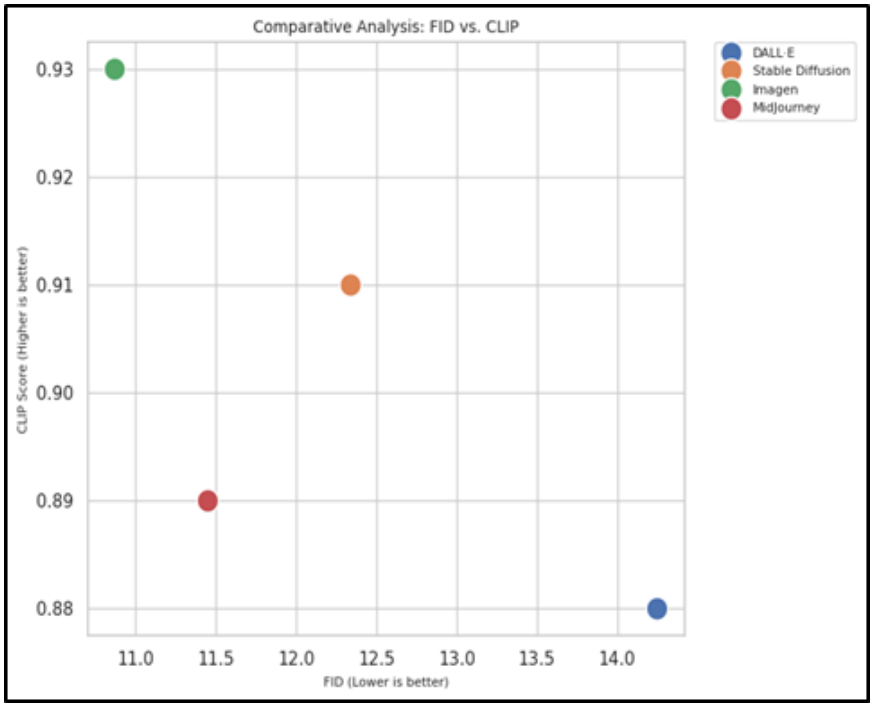

Figure 5. Scatter plot for comparative analysis of FID vs CLIP for all models

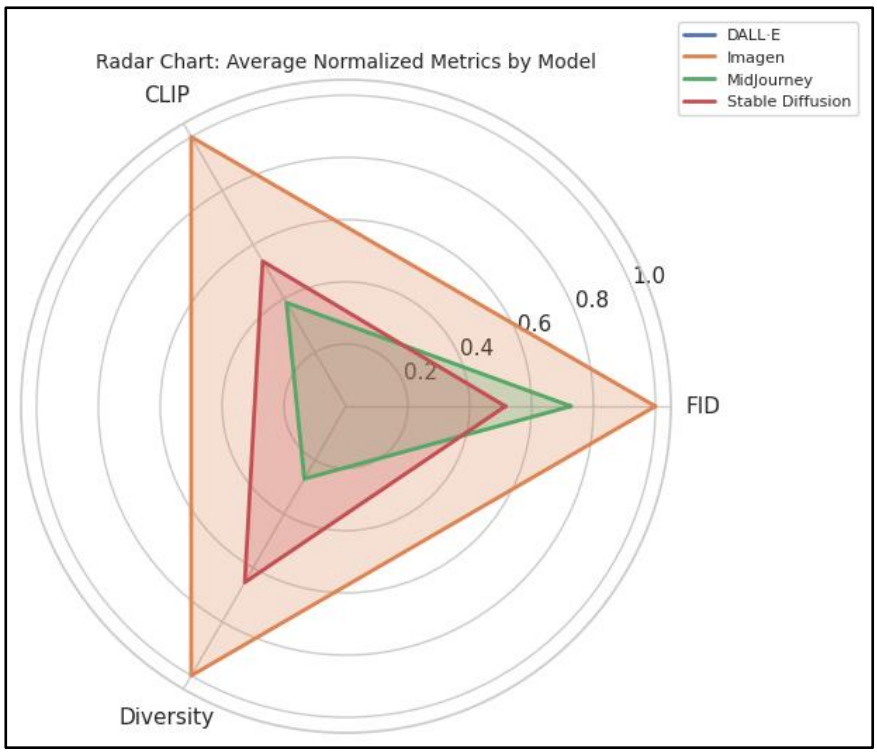

Figure 6. Radar chart for model wise average normalized metrics

Table 11. Model Specific Results analysis

Model	Analysis of FID (↓)	Analysis of CLIP Score (↑)
DALL·E	Struggles to keep up with newer models; generates realistic images but falls short on the fine details needed for a lower FID score.	Aligns well with text, though not exceptionally; effectively translates prompts into visual elements.
Stable Diffusion	Shows notable improvement in fidelity; latent diffusion effectively minimizes noise and boosts realism.	Demonstrates strong text-to-image alignment, thanks to powerful text embeddings such as CLIP.
Imagen	Achieves the top FID score, indicating the highest quality and fidelity in the generated images.	Excels in semantic alignment, achieving the highest score among all models.
MidJourney	Performs strongly in producing visually attractive and artistic images; however, the FID is slightly higher compared to Imagen.	Shows solid text alignment but lags behind Imagen and Stable Diffusion when handling complex or detailed prompts.

Figure 5 shows the scatter plot for comparative analysis of the FID vs CLIP for all image generation model. The FID vs. CLIP plot indicates that Imagen is in a position of having the lowest FID and maximum CLIP score, indicating higher image quality and consistency between text and image. MidJourney and Stable Diffusion have mediocre performance, trading off between consistency and quality, while DALL-E is behind in higher FID and lower CLIP, indicating its relative poor performance.

Figure 6 shows the radar chart indicating Imagen with normalized scores around 0.85 for inverted FID, 0.90 for CLIP, and 0.88 for Diversity, reflecting better balance in terms of quality, alignment, and output diversity. Stable Diffusion and MidJourney follow with moderate scores (≈0.80–0.84), and DALL-E lags with scores around 0.75, reflecting relative underperformance. The detailed model-specific analysis highlights the performance of various text-to-image generation models based on FID and CLIP scores is given in table 11.

To ensure the robustness and generalizability, we employed k-fold cross-validation with on the datasets used in our experiments. In this method, each dataset was partitioned into five approximately equal folds. The model was trained on four folds and tested on the remaining fold, repeating this process five times so that each fold served as the test set once. This procedure

helps mitigate bias from any particular data split and provides a comprehensive evaluation across different subsets. All evaluation metrics, including Frechet Inception Distance (FID), CLIP score, and diversity score, were averaged over the five folds. The reported mean values and standard deviations reflect the model's stability and consistent performance. For larger datasets, where computational cost limited exhaustive cross-validation, we adhered to standard 80-20 train-test splits consistent with prior literature, ensuring fair comparison.

9.3 Model Complexity Comparison

Here, we compare four prominent text-to-image generation models—DALL·E, Stable Diffusion, Imagen, and MidJourney—across key complexity dimensions: model parameter count, floating point operations per second (FLOPS), typical training time, and estimated energy consumption. The parameter count correlates strongly with the resource demands during both training and inference. DALL·E's 12 billion parameters significantly exceed those of Stable Diffusion and Imagen, contributing to longer training times and higher energy use. Stable Diffusion's architecture leverages a latent space to reduce computational overhead, making it comparatively more sustainable for practical applications. Imagen, while smaller in parameters than DALL·E, incurs high FLOPS due to its integrated large language model conditioning. MidJourney's exact complexity metrics remain proprietary; however, estimates suggest a balance between quality and resource demand.

These computational complexities underscore the sustainability trade-offs inherent in text-to-image generation: models achieving state-of-the-art image quality typically require substantial energy and compute resources. This analysis complements the performance metrics detailed earlier in the manuscript and provides a foundation for evaluating environmental impacts and deployment feasibility.

10. Visual Comparison of Generated Images

For good text-to-image creation, proper visual comparison is a must so that one can identify how good models interpret different text prompts and convert them into meaningful pictures. The following are some crucial factors always examined while visually comparing pictures produced by different systems:

a **Creativity:** Creative AI systems are those capable of producing images beyond the literal interpretation of a text prompt; they add unique imaginative or artistic touches which enhance visual output [39]. A perfect example of such models is MidJourney, which produces stylized art exceeding what one would anticipate from a single input phrase— in short, its creativity level is off the charts**.**

Example: A prompt like "a young man crowdsurfing in a movie theatre" could result in various interpretations—one model might focus on realism, while another adds dream-like or abstract elements.

b **Coherence:** Coherence is an essential factor used when measuring the extent to which images align with prompts. Coherent images in this context are ones where every object, person, thing and place in them fits exactly what is described and how it is being described**.**

Example: If the prompt is "a young girl crowdsurfing in a movie theatre," a coherent image would show a girl in a different pose, with no confusing or irrelevant visual elements.

c **Realism:** The realism of a generated image is determined by how much it resembles a real-world object or scene in terms of texture quality, lighting effects applied to it and general environmental characteristics.

Example: For a prompt like "a young man crowdsurfing in a movie theatre," realism would be judged based on how close the generated image looks to an actual photograph in terms of lighting, shadows, and texture.

With respect to above factors, we have generated sample images from the DALL-E2 model, MidJourney and Stable diffusion model. For all the models the given prompt is "a young man crowdsurfing in a movie theatre" and "a young girl crowdsurfing in a movie theatre". Figure 7 shows the images for the prompts generated from the DALL-E2 model with the given prompt. It is observed from the figure 7, images generated from DALL- E2 model are highly creative, contextually accurate, but occasionally over-abstract. Figure 8 shows the sample images generated for the same given prompt to the MidJourney model and the results are quite good as compare to DALL-E2 model. From the generated by MidJourney are more stylized, artistic images, less photorealistic but strong in aesthetics content wise. Figure 9 shows the sample images generated for the same given prompt as "a young man crowdsurfing in a movie theatre" and "a young girl crowdsurfing in a movie theatre" to the Stable Diffusion model and the generated images are high-quality photorealism, but limited on complex scene generation.

Figure 7. Sample images generated from the DALL-E2 model

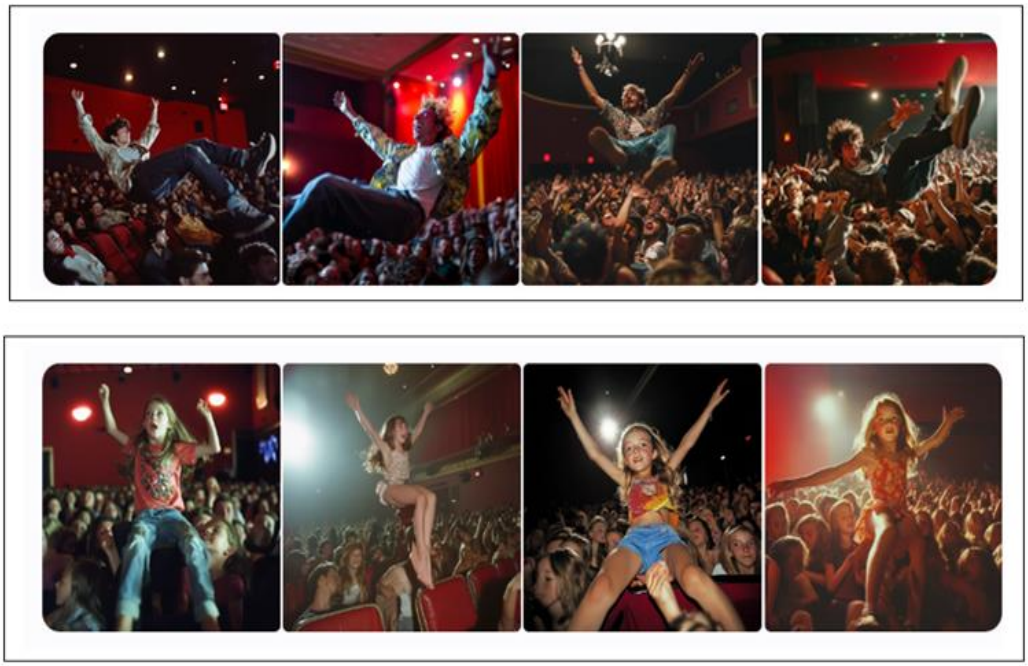

Figure 8. Sample images generated from the MidJourney

Figure 9. Sample images generated from the stable diffusion model

10.1 Trade-offs between Image Quality and Sustainability

One of the critical challenges in the practical deployment of text-to-image generation systems lies in balancing image quality with sustainability considerations, including computational efficiency, environmental impact, and ethical factors. High-fidelity image synthesis typically requires large-scale models and extensive training, which increases energy consumption and carbon footprint. Conversely, optimizing for sustainability can degrade image quality, limiting model applicability. Image quality refers to the visual realism, semantic alignment, diversity, and resolution of generated images. Advanced architectures such as diffusion models and large transformer-based models have pushed state-of-the-art fidelity but require substantial compute resources. Sustainability includes computational cost (training/inference time), energy consumption, carbon emissions, and ethical factors such as fairness and bias mitigation. Models that demand high computational budgets exacerbate environmental concerns and limit deployment in resource-constrained settings. Table 12 and Table 13 shows the qualitative and quantitative trade-offs.

High-quality outputs (low FID) often come at the cost of significantly increased energy usage and carbon footprint, as demonstrated by diffusion and transformer models. This restricts their ability to be environmentally sustainable, particularly in commercial or large-scale applications. With faster inference, moderate energy consumption, and decent image quality, GANs provide a useful compromise. However, their robustness is reduced by mode collapse and training instability. The integration of bias mitigation tends to increase computational overhead, adding another layer to the trade-off, but fairness and ethical considerations are generally under-addressed across models. Application-specific requirements must be taken into account when selecting a model for real-world deployment. For example, high-fidelity artistic generation may justify the costs of diffusion models, while rapid prototyping may prefer GANs. Computational demands can be reduced without compromising the image quality can be done by techniques like pruning, knowledge distillation, and adaptive sampling.

Table 12. Qualitative Model-Specific Trade-offs

Model Type	Image Quality	Sustainability Challenges	Practical Implications	Citations
GANs (e.g., AttnGAN)	Good fidelity and fast inference; prone to artifacts and mode collapse	Moderate training cost and energy use; unstable training impacts efficiency	Suitable for applications requiring fast output with moderate quality; less ideal for ultra-high fidelity needs	[49, 51]
Diffusion Models (e.g., Imagen, Stable Diffusion)	State-of-the-art photorealism and diversity	High computational cost due to iterative denoising steps; large energy consumption	Ideal for quality-critical tasks; high energy cost limits real-time or large-scale deployment without optimization	[14,15,16]
Transformer Models (e.g., DALL·E 3, Parti)	Superior semantic alignment and zero-shot capabilities.	Very high training and inference cost; largest carbon footprint.	Powerful but resource-intensive; challenging to scale sustainably or deploy on edge devices	[16, 17]

Table 13. Quantitative Model Specific Trade-offs

Model Type	Training Cost (GPU-hours)	Inference Latency (s)	Energy Consumption (kWh)	Image Fidelity (FID Score)	Fairness/Bias Mitigation Level	Citations
GAN (AttnGAN)	~1500	<1	Moderate (~200)	25-30	Low	[18]
Diffusion (Imagen)	~3000	~10	High (~600)	8-12	Moderate	[15]
Transformer (DALL·E 3)	>3500	~5	Very High (~800)	10-15	Low to Moderate	[16]

*FID = Frechet Inception Distance (lower is better)

11. Research Gaps

By analysing the in depth literature following research gaps are identified.

- GANs faces the challenges like unstable training and limited bias handling which affect their consistency. The mode collapse tendency also limits the diversity of image generation reducing their sustainability despite the efficient performance.
- Diffusion models have significant environmental problems due to their long inference times and large energy costs, despite producing better images and more reliable training. But they are at early stages and do not have standardized evaluation.
- Transformer architectures push the boundaries of generation diversity and semantic accuracy, yet because of their extensive parameterization, they have the highest environmental costs. Despite minor dataset-level fairness filtering, transformers have yet to adopt comprehensive ethical frameworks, presenting a critical risk area.
- There is a need for multi-objective optimization frameworks that simultaneously address image fidelity, energy efficiency, and fairness.
- The development of standardized sustainability benchmarks, combining carbon footprint, fairness scores, and visual quality, would enable more transparent comparisons.
- Greater emphasis on ethical auditing and bias mitigation pipelines, embedded throughout model development and deployment, is critical.

12. Challenges in Text-to-Image Generation

Text-to-image synthesis has improved a lot thanks to models such as DALL-E, MidJourney, and Stable Diffusion. Nonetheless, there remain a few important challenges and limitations in this field. Some of the challenges and limitations are discussed below [34].

12.1 Semantic coherence

The challenge in generating images is that the generated image ought to accurately represent the semantic meaning of the input text. Often times, models cannot grasp finer details in complex languages and thus are not able to reproduce them correctly.

However simple text prompts tend to yield fairly accurate images, the opposite is also true when it comes to complex descriptions.

12.2 Composition of a Complex Scene

It is very hard to make a coherent and realistic image of more than one object or scenes that are in themselves complex. Coherence of the space, coherence and logical relation between objects, and depth perception are the most troublesome. The main drawback is very often, for current models, they will produce images where elements are misplaced, inconsistent, or do not reflect the complexity of the scene described in the text.

12.3 Fine-Grained Control

Users do not have full control of the exact features they want the final generated image to contain. Although high-level general features can be matched with the input text, tweaking fine details like precise colours, textures, or spatial arrangements is not easy. It is too challenging to fine-tune the output to meet the precise requirements of the user. While current approaches are doing well with overall scenes, they lack control in custom image generation at granular levels.

12.4 Image Resolution and Quality

Generation of images in high resolution is computationally expensive, and retaining this level of detail, especially for larger images, remains elusive. There remains a challenge of quality versus speed balance [38]. Text-to-image models commonly produce blurry details in the high-resolution settings or fail to have photorealistic features, especially fine textures or details of designs.

12.5 Bias and Ethical Considerations

The result of trains from which generative models arise can be biased giving rise to aspects such as gender, racial or cultural bias in the images generated. Also, they often inadvertently reproduce stereotypes. When applied in sensitive industries or on the general public as well as when not carefully watched over they can as well produce undesirable images or texts.

13. Future Research Directions and Potential Improvements

The field of text-to-image generation is rapidly advancing, revealing several exciting research directions and potential enhancements. These areas present opportunities to improve the capabilities of existing models, tackle their limitations, and broaden their applications across different sectors.

13.1 Advancement in Multimodal Models

Future research should prioritize enhancing the integration of multimodal data to better align text and image generation. While current models like DALL-E 2 and Stable Diffusion have made significant strides, they still struggle with understanding complex textual descriptions. To tackle this issue, research could investigate, Creating more advanced cross-modal embeddings that capture subtle relationships between text and images, thereby improving contextual accuracy and coherence. Merging text-to-image generation with other generative tasks, such as text-to-video or 3D model generation, to develop comprehensive multimodal systems that provide richer and more interactive content.

13.2 Improving Model Efficiency and Scalability

The high computational requirements of cutting-edge models often restrict their accessibility and practical application. Future efforts should focus on enhancing model efficiency and scalability by: • Optimized Architectures: Creating more efficient neural network designs that minimize computational demands while preserving or boosting performance. Approaches like pruning, quantization, and knowledge distillation could be investigated. • Resource-Efficient Training: Formulating strategies for training models with reduced resources, possibly leveraging transfer learning or few-shot learning techniques to lessen the reliance on large datasets and significant computational power.

13.3 Addressing Ethical and Bias Concerns

Solving ethical issues, as well as cancelling out bias, becomes critical as text-to-image synthesis models are increasingly being embedded in social applications. The following areas require further research: Anti-bias measures: Usage of sophisticated methods which can easily detect and fix prejudiced elements on the training sets and developed models. Tools might be developed here that will check model behavior thus resulting to its unbiasedness or evenness as it generates items equally and in a fair way.

Ethical standards: Creating specific ethical codes or principles related to good practice of creative machine application that enhance model datasets integrity and clearness in terms of algorithm use.

13.4 Improving Evaluation Metrics

Current evaluation metrics like Inception Score (IS) and Frechet Inception Distance (FID) have their shortcomings when it comes to fully capturing the quality and creativity of generated images. Future research should focus on: establishment of fresh evaluation metrics that more effectively evaluate elements such as artistic value, semantic coherence, and contextual relevance. Including human feedback and subjective evaluations could significantly improve the assessment process. Evaluation frameworks that combine various aspects of image quality and generation effectiveness, taking into account usability, user satisfaction, and alignment with a range of prompts.

14. Discussion towards more Sustainable Text-to-Image Models

Making models more sustainable is essential as text-to-image generation technology advances. In order to strike a balance between innovation and social and environmental responsibility, future models can improve sustainability in a number of important ways:

- Energy-Aware Model Design: creating architectures like lightweight transformers, effective diffusion sampling, and adaptive inference methods that are optimized for reduced energy consumption without sacrificing fidelity.
- Data and Training Efficiency: To cut down on expensive retraining and the need for large amounts of data, self-supervised, few-shot, and transfer learning techniques are used.
- Model compression and acceleration: When pruning, quantization, and knowledge distillation are applied methodically, model size and inference time can be significantly decreased, allowing deployment on devices with limited resources.
- Green AI Infrastructure: promoting the use of data centers powered by renewable energy sources and enhancing hardware efficiency with accelerators designed specifically for AI.
- While fairness and bias are fundamental ethical concerns, a number of other issues demand immediate attention. Some of them are listed below.
- Potential for Misuse and Deep fakes: The capacity to produce incredibly lifelike visuals increases the dangers of false information, fake news, and identity theft, all of which can have significant negative effects on society.
- Intellectual Property (IP) Rights: Since generated images frequently incorporate elements from training data, they present difficult and legally unresolved issues regarding ownership, copyright infringement, and fair use.
- User Consent and Privacy: When training data includes personal or copyrighted images, consent mechanisms and privacy safeguards are necessary to prevent misuse.

- Accountability and Transparency: Models should incorporate mechanisms for provenance tracking and explainability to enable auditability and responsible use.

13. Conclusion

Recent developments in text-to-image generation have brought forward several sophisticated models, including DALL·E 2, Stable Diffusion, MidJourney, and Imagen, marking substantial progress in the ability to convert text into detailed and semantically meaningful images. Our systematic evaluation, using metrics such as FID, CLIP and diversity score demonstrates that Imagen leads in image quality and semantic accuracy, whereas Stable Diffusion shows superior diversity in generated outputs. This survey advances the literature by encompassing comprehensive sustainability framework that encompasses not only computational efficiency and ethical dimensions areas often overlooked in earlier reviews. We provide a critical comparison of major generative architectures, exploring the balance between visual performance and sustainability, and we identify important gaps related to benchmarking standards and fairness.

Nevertheless, challenges remain. These include difficulties in preserving semantic coherence in complex scenes, achieving ultra-high-resolution images, and tackling broader ethical issues such as bias, misuse, and intellectual property rights. The larger computational resources further requires the more resource concise methods.

Future research work will be focus on improving the accuracy while addressing bias mitigation and reducing environmental impact by the models. By evolving the robust ethical methods and unified standards in text to image generation models, enhancement in quality of image generation is possible along with fair, sustainable, and reliable with real world use.

References

[1] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative adversarial nets. Advances in neural information processing systems, 27 (2014).

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners. OpenAI blog, 1(8), (2019) 9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[3] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, Synthetic data augmentation using GAN for improved liver lesion classification. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, Washington, DC, USA, (2018) 289-293. https://doi.org/10.1109/ISBI.2018.8363570

[4] Y. Zhao, H. Wang, J. Zhang, Z. Xu, AI-driven fashion design and customization: Generative adversarial networks in apparel prototyping. Computers in Industry, 127, (2021) 103434.

[5] X. Mao, W. Yu, K.D. Yamada, M.R. Zielewski, Procedural content generation via generative artificial intelligence. arXiv preprint arXiv:2407.09013, (2024). https://doi.org/10.48550/arXiv.2407.09013

[6] S. Lee, M. Kohga, S. Landau, S. O'Modhrain, H. Subramonyam, AltCanvas: a tile-based editor for visual content creation with generative AI for blind or visually impaired people. In Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility, 70, (2024) 1-22. https://doi.org/10.1145/3663548.3675600

[7] C. Zhang, C. Zhang, M. Zhang, I.S. Kweon, (2023). Text-to-image diffusion models in generative AI: A survey. arXiv preprint arXiv:2303.07909. https://doi.org/10.48550/arXiv.2303.07909

[8] 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, 10(4), (2020) e1345.

[9] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for modern deep learning research. In Proceedings of the AAAI conference on artificial intelligence, 34(09), (2020) 13693-13696). https://doi.org/10.1609/aaai.v34i09.7123

[10] P. Cao, F. Zhou, Q. Song, L. Yang, (2024) Controllable generation with text-to-image diffusion models: A survey. arXiv preprint arXiv:2403.04279. https://doi.org/10.48550/arXiv.2403.04279

[11] R. Schwartz, J. Dodge, N.A. Smith, O. Etzioni, Green AI. Communications of the ACM, 63(12), (2020) 54-63. https://doi.org/10.1145/3381831

[12] A. Birhane, V.U. Prabhu, E. Kahembwe, (2021) Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963. https://doi.org/10.48550/arXiv.2110.01963

[13] I.O. Gallegos, R.A. Rossi, J. Barrow, M.M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N.K. Ahmed,. Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), (2024) 1097-1179. https://doi.org/10.1162/coli_a_00524

[14] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.M. Munguia, D. Rothchild, D. So, M. Texier, J. Dean, (2021).Carbon emissions and large neural

network training. arXiv preprint arXiv:2104.10350. https://doi.org/10.48550/arXiv.2104.10350

[15] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E.L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35, (2022) 36479-36494. https://doi.org/10.48550/arXiv.2205.11487

[16] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation. In International conference on machine learning, (2021) 8821-8831. https://doi.org/10.48550/arXiv.2102.12092

[17] E.M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, (2021). On the dangers of stochastic parrots: Can language models be too big? FAccT 2021. https://doi.org/10.1145/3442188.3445922

[18] Y. Gong, Z. Zhan, Q. Jin, Y. Li, Y. Idelbayev, X. Liu, A. Zharkov, K. Aberman, S. Tulyakov, Y. Wang, J. Ren, (2024). Efficient GANs for Image-to-Image Translation. arXiv preprint arXiv:2401.06127.

[19] A. Valyaeva, AI image statistics: How much content was created by ai. Everypixel Journal, 15, (2023). Adobe Creative Cloud Adoption Grows to 33 Million Paid Members, https://prodesigntools.com/number-of-creative-cloud-subscribers.html

[20] M. Yang, Z. Wang. Image synthesis under limited data: A survey and taxonomy. International Journal of Computer Vision, (2025) 1-38.

[21] H. Chen, Q. Xiang, J. Hu, M. Ye, C. Yu, H. Cheng, L. Zhang. Comprehensive exploration of diffusion models in image generation: a survey. Artificial Intelligence Review, 58(4), (2025) 99. https://doi.org/10.1007/s10462-025-11110-3

[22] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee. Generative adversarial text to image synthesis. In International conference on machine learning, PMLR, (2016) 1060-1069.

[23] A. Radford, L. Metz, S. Chintala, (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv: 1511.06434.

[24] D.P. Kingma, M. Welling, An introduction to variational autoencoders. Foundations and Trends® in Machine Learning, 12(4), (2019) 307–392. http://dx.doi.org/10.1561/2200000056

[25] V. Ashish, Attention is all you need. Advances in neural information processing systems, 30, (2017).

[26] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, Long Ouyang et al. Improving image generation with better captions. Computer Science, 2(3), (2023) 8.

[27] J. Bao, W.M. Yu, K. Yang, C. Liu, T.J. Cui. Improved few-shot SAR image generation by enhancing diversity. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,IEEE, 17, (2024) 3394-3408. https://doi.org/10.1109/JSTARS.2024.3352237

[28] J. Ho, A. Jain, P. Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, (2020) 6840-6851.

[29] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision, IEEE, Venice, Italy, (2017) 5907-5915. https://doi.org/10.1109/ICCV.2017.629

[30] T. Xu, H. Zhang, X. Huang, S. Zhang, L. Zhang, (2018). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1316-1324.

[31] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2204.06125, 1(2), (2022) 3.

[32] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew,

[33] I. Sutskever, M. Chen. (2021) Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.

[34] E. D'Armenio, A. Deliège, M.G. Dondero. Semiotics of Machinic Co-Enunciation. About Generative Models (Midjourney and DALL· E), Signata. Annals of Semiotics, 15, (2024). https://doi.org/10.4000/127x4

[35] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence, 41(8), (2018) 1947-1962. https://doi.org/10.1109/TPAMI.2018.2856256

[36] B. Li, X. Qi, T. Lukasiewicz, P.H.S. Torr, ManiGAN: Text-guided image manipulation, in Proc. IEEE/CVF Conference Computer Vision Pattern Recognition. (CVPR), IEEE, Seattle, WA, USA, (2020) 7877–7886. https://doi.org/10.1109/CVPR42600.2020.00790

[37] M.A.H. Palash, M.A. Al Nasim, A. Dhali, F. Afrin. Fine-grained image generation from bangla text description using attentional generative adversarial network. In 2021 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON),

IEEE, Dhaka, Bangladesh, (2021) 79-84. https://doi.org/10.1109/RAAICON54709.2021.9929536

[38] M. Bahani, A. El Ouaazizi, K. Maalmi. The effectiveness of T5, GPT-2, and BERT on text-to-image generation task. Pattern Recognition Letters, 173 (2023) 57-63. https://doi.org/10.1016/j.patrec.2023.08.001

[39] M. Kang, J.Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, T. Park. Scaling up GAN's for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, BC, Canada, (2023)10124-10134. https://doi.org/10.1109/CVPR52729.2023.00976

[40] A. Sanghi, H. Chu, J.G. Lambourne, Y. Wang, C.Y. Cheng, M. Fumero, K.R. Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA, (2022) 18603-18613. https://doi.org/10.1109/CVPR52688.2022.01805

[41] J. Yu, Y. Xu, J.Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B.K. Ayan, B. Hutchinson, Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 2(3), (2022) 5.

[42] A. Sanghi, H. Chu, J. G. Lambourne, Y.Wang, C.Y. Cheng, M. Fumero K.R. Malekshan, (2021) CLIP-Forge: Towards zero-shot text-to-shape generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, USA. https://doi.org/10.1109/CVPR52688.2022.01805

[43] Z. Wang, W. Liu, Q. He, X. Wu, Z. Yi, (2022) Clip-gen: Language-free training of a text-to-image generator with clip. arXiv preprint arXiv:2203.00386. https://doi.org/10.48550/arXiv.2203.00386

[44] J. Shi, C. Wu, J. Liang, X. Liu, N. Duan, DiVAE: Photorealistic images synthesis with denoising diffusion decoder. (2022) arXiv:2206.00386v1.

[45] S. Naveen, M.S.R. Kiran, M. Indupriya, T.V. Manikanta, P.V. Sudeep, Transformer models for enhancing AttnGAN based text to image generation. Image and Vision Computing, 115, (2021) 104284. https://doi.org/10.1016/j.imavis.2021.104284

[46] H. Bansal, D. Yin, M. Monajatipoor, K.W. Chang, (2022) How well can text-to-image generative models understand ethical natural language interventions?. arXiv preprint arXiv:2210.15230. https://doi.org/10.18653/v1/2022.emnlp-main.88

[47] R. Navigli, S. Conia, B. Ross, Biases in large language models: origins, inventory, and discussion. ACM Journal of Data and Information Quality, 15(2), (2023) 1-21. https://doi.org/10.1145/3597307

[48] C. Bird, E. Ungless, A. Kasirzadeh, (2023) August. Typology of risks of generative text-to-image models. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 396-410. https://doi.org/10.1145/3600211.3604722

[49] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J.S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente, M. Kang, Holistic evaluation of text-to-image models. Advances in Neural Information Processing Systems, 36, (2023) 69981-70011.

[50] H. Chen, Q. Xiang, J. Hu, M. Ye, C. Yu, H. Cheng, L. Zhang, Comprehensive exploration of diffusion models in image generation: a survey. Artificial Intelligence Review, 58(4), (2025) 99. https://doi.org/10.1007/s10462-025-11110-3

[51] X. Tu, Z. He, Y. Huang, Z.H. Zhang, M. Yang, J. Zhao, An overview of large AI models and their applications. Visual Intelligence, 2(1), (2024) 34. https://doi.org/10.1007/s44267-024-00065-8

[52] P.P. Liang, C. Wu, L.P. Morency, R. Salakhutdinov, (2021) Towards understanding and mitigating social biases in language models. In International conference on machine learning, PMLR, 139, 6565-6576.

[53] O. Bendel, Image synthesis from an ethical perspective. AI & Soc, 40, (2025) 437–446. https://doi.org/10.1007/s00146-023-01780-4

[54] E. Sheng, K.W. Chang, P. Natarajan, N. Peng, (2019) The woman worked as a babysitter: On biases in language generation. arXiv preprint arXiv:1909.01326. https://doi.org/10.18653/v1/D19-1339

[55] M. Yang, Z. Wang, Image synthesis under limited data: A survey and taxonomy. International Journal of Computer Vision, 133(6), (2025) 3689-3726. https://doi.org/10.1007/s11263-025-02357-y

[56] S. Lin, R. Ji, C. Yan, B. Zhang, L. Cao, Q. Ye, F. Huang, D. Doermann, (2019) Towards optimal structured cnn pruning via generative adversarial learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, IEEE, USA. https://doi.org/10.1109/CVPR.2019.00290

[57] C. Tao, L. Hou, W. Zhang, L. Shang, X. Jiang, Q. Liu, P. Luo, N. Wong, (2022) Compression of generative pre-trained language models via quantization. arXiv preprint arXiv:2203.10705. https://doi.org/10.18653/v1/2022.acl-long.331

[58] F. Zeng, W. Gan, Y. Wang, P.S. Yu, (2023) December. Distributed training of large language models. In 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS), IEEE, China. https://doi.org/10.1109/ICPADS60453.2023.00126

[59] X. Wang, Z. Tang, J. Guo, T. Meng, C. Wang, T. Wang, W. Jia, Empowering edge intelligence: A comprehensive survey on on-device ai models. ACM Computing Surveys, 57(9), (2025) 1-39. https://doi.org/10.1145/3724420

[60] J. Xu, Z. Li, W. Chen, Q. Wang, X. Gao, Q. Cai, Z. Ling, (2024) On-device language models: A comprehensive review. arXiv preprint arXiv:2409.00088. https://doi.org/10.48550/arXiv.2409.00088

[61] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, (2017) GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems 30 (NIPS 2017).

[62] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, X. Chen, (2016) Improved techniques for training GANs. Advances in Neural Information Processing Systems 29 USA: Curran Associates, 2234–2242.

[63] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision. Proceedings of Machine Learning Research, 139, (2021) 8748–8763.

[64] Poloclub, “GitHub - poloclub/diffusiondb: A large-scale text-to-image prompt gallery dataset based on Stable Diffusion,” GitHub. https://github.com/poloclub/diffusiondb

[65] “poloclub/diffusiondb · Datasets at Hugging Face,” Mar. 16, 2023. https://huggingface.co/datasets/poloclub/diffusiondb

[66] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, C. Pal, (2018) Fashion-gen: The generative fashion dataset and challenge. arXiv preprint arXiv:1806.08317.

[67] Cyizhuo, “GitHub - cyizhuo/CUB_200_2011_dataset: CUB-200-2011 dataset by classes folder,” GitHub. https://github.com/cyizhuo/CUB_200_2011_dataset

Authors Contribution Statement

Smita Bharne: Conceptualization, Methodology, Experimental Work, Data Curation, Writing - Original Draft. Pallavi Sapkale: Investigation, Review & Editing. Ekta Sarda: Formal Analysis, Visualization, Supervision. Puja Padiya: Review & Editing, Validation. Shamal Salunkhe: Review & Editing, Resources. All the authors read and approved the final version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License