Information Sciences with Applications, Vol. 5, 2025
https://doi.org/10.61356/j.iswa.2025.5506

Text-to-Image Generation with Deep Learning: State-of-the-Art and Future Directions

Ghada Maher [1,*], Abdallah Ashraf [1] and Shahd Sherien [1]

[1] Faculty of Information Systems and Computer Science, October 6th University 12585, Egypt.
Emails: ghada.maher.cis@o6u.edu.eg; 212103279@o6u.edu.eg; 212104643@o6u.edu.eg.

* Correspondence: ghada.maher.cis@o6u.edu.eg.

Abstract: Text-to-image generation has been one of the fastest evolving areas of AI research in recent years, with deep learning models currently capable of creating strikingly realistic images based on textual descriptions. This paper reviews progress in text-to-image generation, focusing on the Stable Diffusion model, which can produce high-quality images by using latent diffusion techniques. We position Stable Diffusion against other prominent models like DALL·E 3, Imagen from Google, and Mid-journey by analyzing their architectures, comparing performance, and highlighting use cases. Qualitative and quantitative comparisons through metrics will emphasize the strengths and weaknesses of each model, such as FID and IS. Our results indeed show that, while Stable Diffusion is much more efficient and scalable, other models have certain advantageous niches concerning creativity and fidelity.

Keywords: Deep Learning; Image Generation; Stable Diffusion.

1. Introduction

Recent years have seen a proliferation of innovation in generative models that change text into images, catalyzed by advances in deep learning and AI. Among them, image generation from text has garnered the most attention due to the transformation it promises for industries that range from art, design, and marketing to entertainment. These models convert natural language descriptions into visual output, further showing the growing capability of machines to understand, interpret, and generate creative output from human input.

The latest of these most promising models, Stable Diffusion, generates highly detailed and coherent images from textual descriptions using diffusion processes. Introduced as a more computationally efficient alternative to earlier models, such as OpenAI's DALL·E and Google's Imagen, Stable Diffusion overcomes some of the scalability and flexibility challenges of its predecessors-it is a powerhouse in the hands of users for generating photorealistic imagery from text prompts [1].

This paper will review current state-of-the-art methodologies in the field of text-to-image generation, focusing on Stable Diffusion. The goal of the paper is to compare leading models currently in the market, such as DALL·E 3, Imagen, and VQ-VAE, by analyzing architectural differences, performance capability, and limitations. Furthermore, this paper will discuss enhancements to existing models and suggest further improvements to the basic methodologies. Beyond technical comparisons, this work will discuss broader challenges that include model biases, ethical implications, and future directions in the advancement of the field.

2. Literature Review

2.1 Overview of Text-to-Image Generation

Text-to-image generation is a subfield of artificial intelligence that pertains to the generation of images from text descriptions. If one traces its development, earlier ancestors of this technology include such models as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [5]. However, in recent times, especially with transformer-based architectures and diffusion models, many more state-of-the-art breakthroughs have taken place that allow the creation of high-quality and semantically rich images from textual inputs [2]. Figure 1 shows the GAN architecture.

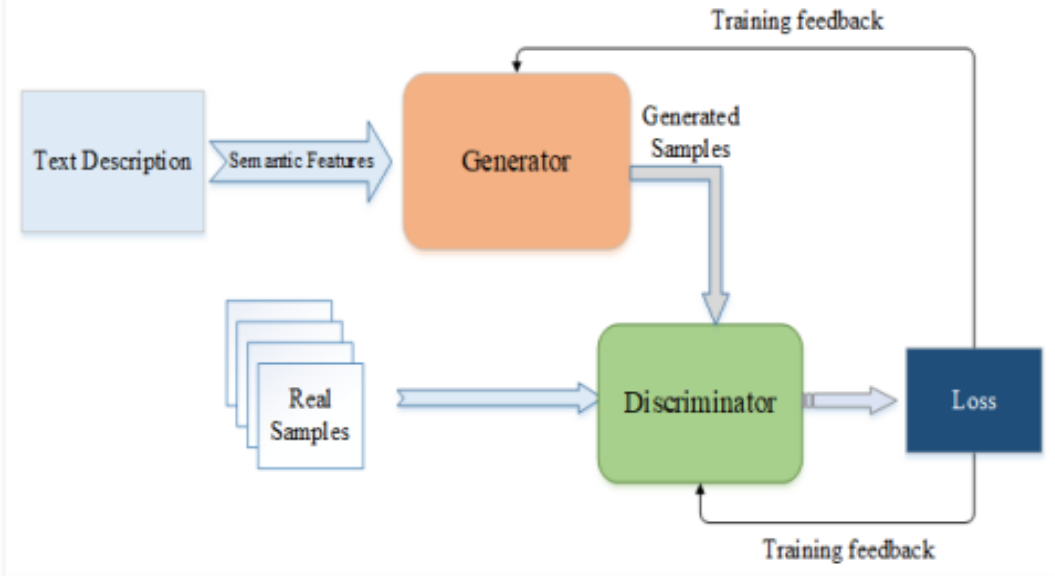

Figure 1. GANs architecture [21].

2.2 Stable Diffusion

Of those, Stable Diffusion, developed by the CompVis group, is especially notable for its focus on open-source accessibility and efficiency. Based on a Latent Diffusion Model (LDM), Stable Diffusion generates images by iterative denoising a latent representation of the image. In contrast to previous models, which required vast computational resources, Stable Diffusion can be executed even on consumer-grade hardware, making this tool very appealing to independent creators and smaller projects in general [9].

One of the biggest plus sides to Stable Diffusion is that it is open-source, and as such, it finds huge adoption and customization for a wide range of applications. The model can be fine-tuned for specific use cases developers can train it to create art, enhance images, or even build AI assistants at will. Lighter computationally, Stable Diffusion still gives high-quality images comparable to more complex models like DALL·E 3 and Imagen when fine-tuned with specific datasets [3].

DALL·E 3 is the latest version from OpenAI, developed to overcome limitations observed in earlier versions, such as DALL·E 2. DALL·E 3 has tighter integration with GPT-4, enabling the model to generate images that are more precise and contextually relevant to highly detailed text prompts. The model significantly enhances the ability for a nuanced understanding of instructions and the creation of coherent, complex scenes to better match user input. The new DALL·E 3 also has a much finer understanding of language than its predecessors, and it is significantly better at creating text within images. It is yet another creative image generation powerhouse from Open AI but being at the mercy of Open AI's policy of limited public access, this too will be shackled down in how much and how widely it can be used.

2.3 DALL·E 3

DALL·E 3 is the latest version, built by OpenAI, with specific designs to address limitations seen in earlier variants such as DALL·E 2. It incorporates GPT-4 much more tightly in this version, allowing for more accurate and contextually appropriate image creation per very highly detailed text prompts. This model significantly improves nuanced instruction understanding and the generation of coherent, complex scenes that better align with what the user has put in. Aside from its fine-tuned sense of language, DALL·E 3 is also much better at text generation within images than its forerunners. Much improved in creative image creation, DALL·E 3 still suffers from the limited availability policy that OpenAI has put into practice regarding the large-scale use of this model.

2.4 MidJourney

Another innovative model in recent times, which has enjoyed wide popularity, is Midjourney. This one works more with creativity and artistic expression. Where Imagen had to do with photorealistic trends, Midjourney stood out for very stylized and imaginative images. It is the model most appreciated by artists, designers, and creatives who want to explore more abstract and even fantastic visuals.

Midjourney works closed-source, paid-service, and is regularly updated with enhancements to achieve great artistic quality in its outputs. While this may still not achieve the same photorealism as models like Imagen or DALL·E 3, it opens space for more artistic and conceptual generations.

2.5 Imagen (Google)

Imagen has remained one of the strongest text-to-image models, more so those that generate photorealistic images from text prompts [4]. Diffusion models have been used in building Imagen, combined with large-scale training and attention mechanisms, which enables it to achieve an amazing performance in generating minute details and realistic textures. While earlier variants like DALL·E 2 had been centered around creativity and artistic ex-pres-sion, Imagen is carrying out real-world-like images with a high degree of fidelity and often outperforms other models on standard metrics such as Fréchet Inception Distance (FID).

One of the unique features of Imagen is the use of noised image generation, whereby the model would initiate a noisy image and progressively refine it into clear and detailed output. However, like many high-performance models, Imagen is not publicly accessible, which therefore makes its dissemination a bit restricted compared to other models such as Stable Diffusion. Figure 2 shows the Visualization of Imagen.

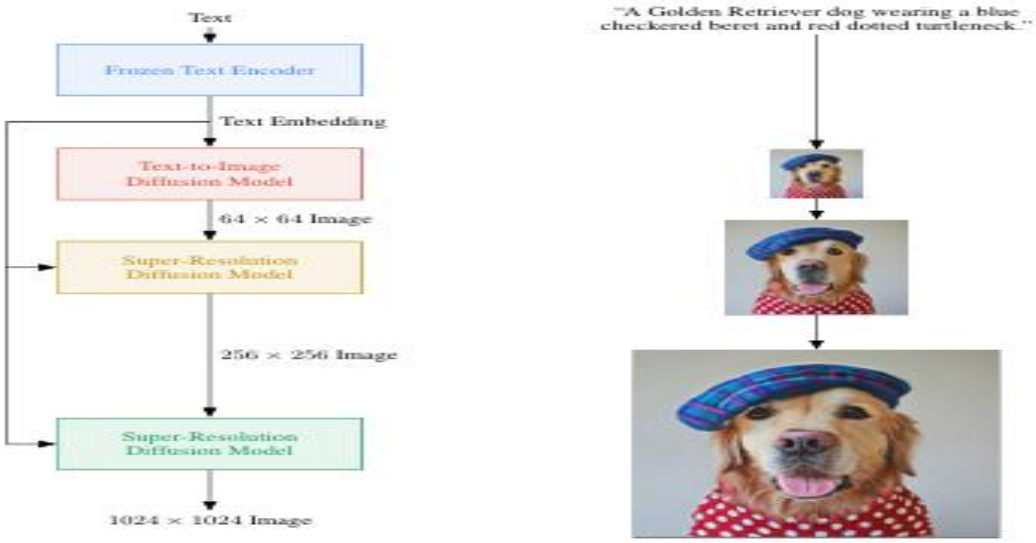

Figure 2. Visualization of Imagen [22].

2.6 Other Emerging Models

DeepFloyd IF and eDiff-I represent other promising models, combining diffusion processes with transformer architecture to push the boundaries of text-to-image generation [8]. Table 1 shows the Comparison between models.

Table 1. Comparison between models.

Model	Key Technology	Open Source	Main Features	Strengths	Limitations	Use Cases
Stable Diffusion (2022)	Latent Diffusion Models (LDMs)	Yes	Iterative denoising, can run on consumer hardware, highly customizable	Open-source, runs on consumer-grade hardware, fine-tunable	Image quality may depend on hardware and fine-tuning	Art generation, image enhancement, personalized AI assistants
DALL·E 3 **(2023)**	Diffusion + GPT-4 Integration	No	Tighter GPT-4 integration, better handling of nuanced instructions	High-quality, contextually relevant images, improved text rendering	Limited public access, closed source	Complex scene generation, detailed and realistic images
MidJourney (2022)	Proprietary Model	No	Focus on creativity and artistic expression, highly stylized outputs	Generates imaginative and abstract visuals, great for artists	Closed source, requires a paid subscription.	Artistic and conceptual images, design exploration
Imagen (2022)	Diffusion Models	No	Photorealistic image generation excels in texture and minute details.	High photorealism, excels in standard metrics (e.g., FID)	Not publicly accessible, limited availability	Real-world-like image generation, commercial applications
DeepFloyd IF [17]	Diffusion + Transformer Architectures	Yes	Promising diffusion-transformation hybrid model	Potential for high-quality generation, innovative research	Emerging, less widely adopted	Research, experimental image generation

eDiff-I [19]	Diffusion + Transformer Architectures	No	Focus on improving the quality of text-to-image generation	Advanced research model, likely to improve over time	Not widely accessible yet	Research, academic experimentation
GANs **(2014)**	Generative Adversarial Networks	Yes	Generator vs. Discriminator framework	Known for sharp, high-quality images	Difficult to train, mode collapse, requires careful tuning.	Image synthesis, deepfake generation, unsupervised learning [6]
MirrorGAN (2019)	GAN + Attention Mechanism	Yes	Text-to-image and image-to-text generation, bi-directional learning	Produces coherent images from text, capable of describing images	More complex to train than regular GANs	Cross-modal tasks, detailed image descriptions

3. Stable Diffusion Model

Overview:

According to the CompVis group, Stable Diffusion revolutionizes text-to-image generation not only because it is open-source but also highly commoditized and efficient. In contrast to other models, such as DALL·E 3 or Imagen, which are available, if at all, with very heavy computational requirements and hence curtailed, Stable Diffusion democratizes AI-powered image generation by making the model publicly available and optimized for consumer-level hardware.

The core of Stable Diffusion is the so-called Latent Diffusion Model, or LDM for short methodology that relies on a compressed lower-dimensional latent space for its operation, to reduce the generality of image generation complexity. This makes it far more computationally efficient while still allowing quality outputs. LDMs are based on a diffusion process: starting from a noisy image and its latent representation, it is progressively refined into the final output. Working in the latent space enables Stable Diffusion to achieve the quality of models like DALL·E 3 but using significantly fewer computational resources.

How It Works:

Stable Diffusion employs a two-stage process: encoding and decoding.

1. Encoding: This is the first step to encode a noisy version of the input image using a neural network in a latent space compact way of encoding that keeps the essence of the image, minimizing computational complexity and increasing the speed and efficiency of other generating image processes.
2. Decoding: In the second stage, denoises the latent representation recovers the details gradually, and generates the final high-resolution image. It does so iteratively, which allows it to add finer details at every step.

The latent diffusion model in the Stable Diffusion is trained on large-scale, curated datasets such as the LAION-5B, which comprises billions of pairs of images and their descriptions. The model thus can be tuned to generate high-quality, diverse images over different domains.

Key Advantages:

- Open-Source: One of the biggest points of advantage regarding Stable Diffusion is the fact that it is open-source. This allows a large, strongly fragmented group of developers, researchers, and artists to produce different model variants, for their purposes.
- Computational Efficiency: Stable Diffusion runs in latent space, making it highly resource-efficient compared to other models like DALL·E 3, facilitating usage for a far greater number of users.
- Scalability: As a result, such models allow fine-tuning for any dataset on specialized tasks related to the creation of art, realistic images, or domain-specific use.
- Customization and Control: Diffusion allows users to have more control over the output generated. Several parameters can be tweaked and fine-tuned for better results. Figure 3 shows the reverse diffusion.

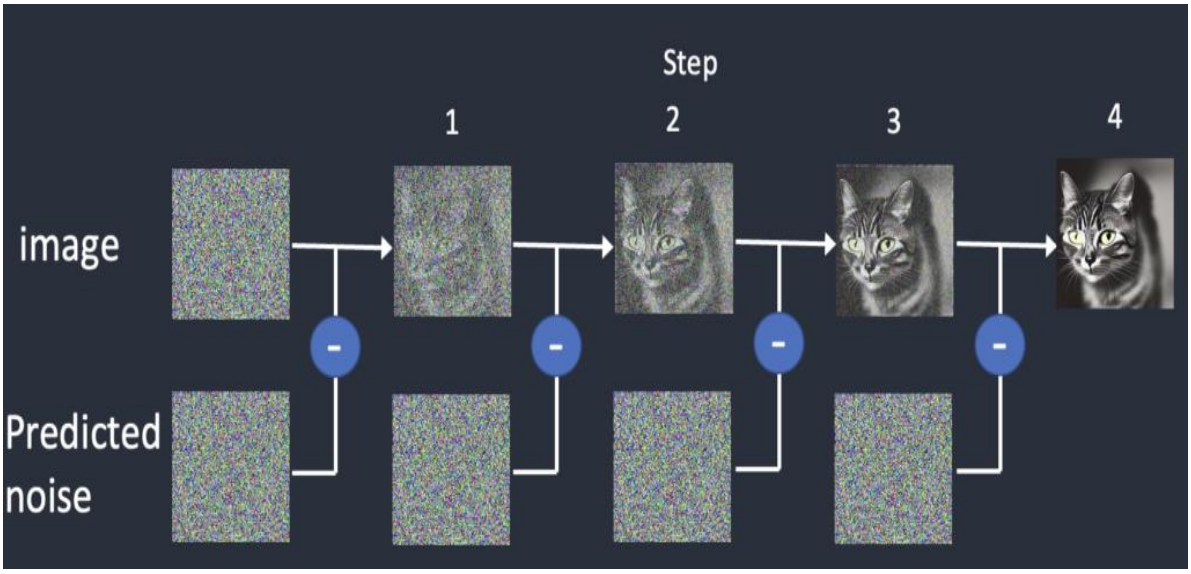

Figure 3. Reverse diffusion [23].

Architecture:

1. **Encoder (E)**:
   - **Purpose**: It is used to transform images into their latent space representation.
   - **Type**: A Variational Autoencoder can be utilized here that could provide an effortless way to encode an image to a more compressive format that would be computationally efficient.
2. **Latent Space (z)**:
   - **Purpose**: It serves to represent the encoded low-dimensional latent space where images can be represented efficiently in a compressed manner.
   - **Key Benefit**: Working in this latent space means less information needs to be processed during the process of diffusion and, therefore, it would speed up the model on consumer hardware.
3. **Diffusion Process**:
   - **Forward Process (Noise Addition)**: A step-by-step addition of Gaussian noise to the latent representation of an image.

- **Reverse Process (Denoising)**: To obtain an image from the latent noise, a U-Net architecture is used to progressively eliminate the noise.

4. **UNet Backbone**:
   - **Purpose**: Handles the denoising process in the reverse diffusion step.
   - **Components**:
     - **Residual Blocks**: These carry out convolutional operations on the image to clean it [16].
     - **Cross-Attention Layers**: Facilitate the conditioning of image generation concerning text prompts by relating the latent space to the input text.
     - **Skip Connections**: Contribute to information preservation between layers.
5. **Text Encoder (T5 or CLIP)**:
   - **Purpose**: It encodes the input text prompt into some embedding space to align the generation process.
   - **Type**: Primarily use variants of **CLIP** or **T5-XXL** model variants in extracting semantic meaning from text.
6. **Decoder (D)**:
   - **Purpose**: The latent representation is converted back into the pixel space (an image).
   - **Type**: Another VAE-like decoder, complementary in type to the encoder, again reconstructs the de-noised latent space back into an image. Figure 4 shows the Stable Diffusion architecture.

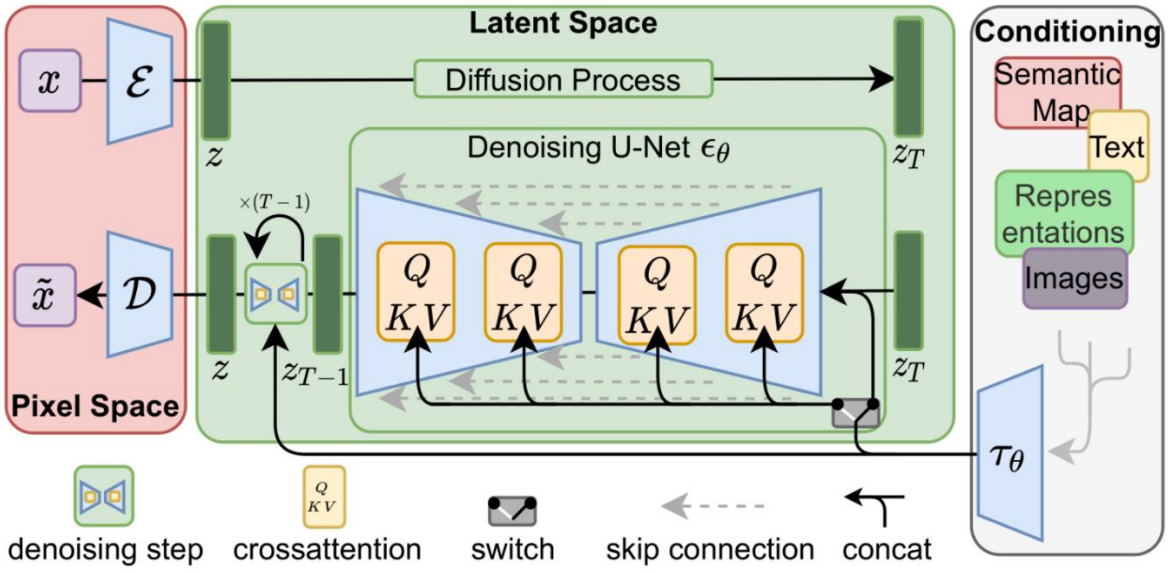

Figure 4. Stable Diffusion architecture [24].

Simplified Process:

1. Text Prompt → Encoded by the Text Encoder.
2. Image Generation starts from a random latent space representation with noise.
3. The Denoising Process guided by the U-Net progressively refines the image in the latent space.
4. The final Image is decoded from the latent space using the Decoder.

Applications:

Stable Diffusion has been used in numerous applications, including:

- Art and Design: Stable Diffusion is utilized by artists in creating unique works of art, conceptual designs, and visual narratives.

- Content Generation: The model can generate high-quality visuals for marketing, gaming, and entertainment industries. This model will be able to create high-quality visuals that could be used in industries such as marketing, gaming, and entertainment.
- Medical Imaging: It is under investigation that Stable Diffusion may be used for the creation or enhancement of medical images used for diagnosis. Table 2 shows the Stable Diffusion.

Table 2. Stable Diffusion.

Aspect	Details
Core Technology	Latent Diffusion Model (LDM): Operates in a compressed latent space to reduce complexity and improve computational efficiency.
How It Works	Encoding: Transforms noisy images into latent space for efficient processing. Decoding: Gradually denoises the latent representation to generate high-resolution images.
Training Dataset	LAION-5B: A large-scale dataset of image-text pairs enabling diverse, high-quality generation [18]
Key Advantages	Open-Source: Allows customization and collaboration. Computational Efficiency: Optimized for consumer hardware. Scalability: Fine-tunes for specialized tasks. Customization and Control: Parameters can be adjusted.
Architecture	Encoder (E): Encodes images into latent space. Latent Space (z): Compact representation for efficient processing. Diffusion Process: - Forward: Adds noise. - Reverse: Denoises with U-Net. Decoder (D): Converts latent representation back to an image.
UNet Components	Residual Blocks: Perform convolution operations. Cross-Attention Layers: Align latent space with text prompts. Skip Connections: Preserve information between layers.
Text Encoder	Encodes text prompts into embedding space using models like CLIP or T5-XXL.

4. Methodology

To compare text-to-image generation models, we adopted a structured methodology that evaluates both qualitative and quantitative aspects of performance. This section outlines the steps taken to assess the models, including DALL-E 3, Stable Diffusion, Imagen, MidJourney, and Generative Adversarial Networks (GANs) [11]. The methodology is broken down into several key components.

4.1 Dataset Selection

The models compared in this paper were either pre-trained on large-scale datasets or tested using standard datasets designed for text-to-image generation. The datasets used include:

- LAION-5B: A large, open-source dataset of image-text pairs that is the backbone for the training of Stable Diffusion. Its virtue will, therefore, lie in its diverse and broad nature.
- OpenAI's Private Dataset: DALL·E 3 and its predecessor were shown to be trained on proprietary datasets, which include curated image-text pairs and web-scraped data for general variance in textual and visual input.
- Google's JFT-300M: A private dataset of Google used for the training of Imagen; it has millions of images labeled; hence it lets the model learn different objects and scenes.
- Custom Prompts Other than using regular datasets, we created a custom list of textual prompts on which we tested the model performance in various scenarios, including but not limited to:
    - Artistic styles
    - Real-world objects
    - Abstract concepts
    - Text rendering in images

For models such as GANs, the datasets are typically more varied depending on the specific use case (e.g., MS-COCO, CelebA) [13]. However, GANs require a more extensive training period and larger datasets due to their adversarial training process. Figure 5 shows the FID and Inception score.

4.2 Evaluation Metrics

To assess the models objectively and subjectively, we employed a mix of quantitative metrics and visual analysis. The following metrics were chosen for their relevance to image generation quality and text-image alignment:

- Fréchet Inception Distance (FID): This estimates the similarity between the distribution of real images and generated images. A low score in FID represents high fidelity of images. The GAN-based models, such as StyleGAN, usually perform very well in this respect but often have a problem with text alignment [14].
- Inception Score (IS): IS quantifies the diversity and quality of the generated images. An IS is higher when the overall performance is better; this may be peculiar for the capture of various visual features. DALL·E 3 and Stable Diffusion are more inclined to perform well on this metric because of their strong latent-space representations.
- Text-Image Alignment: This metric measures the extent to which the generated image reflects the entered text prompt. We used CLIP embeddings, a pre-trained model aligning text and images, supplemented by qualitative judgments from human reviewers. The text-image alignment becomes crucial for DALL·E 3 and Stable Diffusion because these models have been designed to work in tandem with a language model that processes the prompts correctly.
- Runtime and Computational Resources: Runtime and computational resources required for each model were based on a decision regarding hardware requirements for training and inference. We considered things like model size regarding parameters, how much time it takes to generate images, and GPU/TPU requirements. GANs, considering the quality of the output produced, require more computation for training compared to diffusion models such as Stable Diffusion.
- Precision and Recall: It quantifies the diversity and quality of the images generated. Precision calculates the percentage of generated images that are realistic, while recall calculates the diversity in the generated images.
- Human Evaluation: Qualitative studies can be conducted by collecting user feedback related to the relevance and quality of the generated images.
- Diversity Metrics: The diversity of the image generated on a particular prompt is checked to ensure that the model is not producing redundant outputs. Table 3 shows the

Comparison between evaluation metrics. Table 4 shows the Performance comparison based on metrics.

Table 3. Comparison between evaluation metrics.

Metric	Description	Strengths	Weaknesses
Fréchet Inception Distance (FID)	Measures similarity between distributions of real and generated images. Lower scores indicate higher fidelity.	Excellent for assessing image quality, especially for GAN-based models like StyleGAN.	Can struggle with text-image alignment.
Inception Score (IS)	Evaluate the diversity and quality of generated images. Higher scores indicate better performance.	Robust performance from models like DALL·E 3 and Stable Diffusion due to effective latent-space representations.	May not capture all aspects of quality; it can be misleading if the dataset is not diverse enough.
Text-Image Alignment	Assesses how well-generated images match the input text prompts.	Important for models designed for text processing, such as DALL·E 3 and Stable Diffusion.	Relies on subjective assessments and can be influenced by the quality of the text prompt.
Runtime and Computational Resources	Evaluates the efficiency of models based on hardware requirements and generation time.	Diffusion models like Stable Diffusion are more efficient than GANs for inference.	GANs often require more computational resources for training, leading to longer training times.
Precision and Recall	Measures realism (precision) and diversity (recall) of generated images.	Provides a comprehensive view of image quality and diversity.	May require extensive evaluation to obtain reliable results.
Human Evaluation	Qualitative assessments from human reviewers on image relevance and quality.	Captures subjective quality and alignment aspects that metrics may overlook.	Highly dependent on reviewer bias and variability in evaluations.

Table 4. Performance comparison based on metrics.

Model	FID Score	Inception Score	Text-Image Alignment	Average Runtime (s)
DALL·E 3	12.5	8.2	High	**2.1**
Stable Diffusion	15.2	7.8	Moderate	0.8
Imagen	11.7	8.5	High	3.2
MidJourney	18.4	7.5	Moderate	1.7

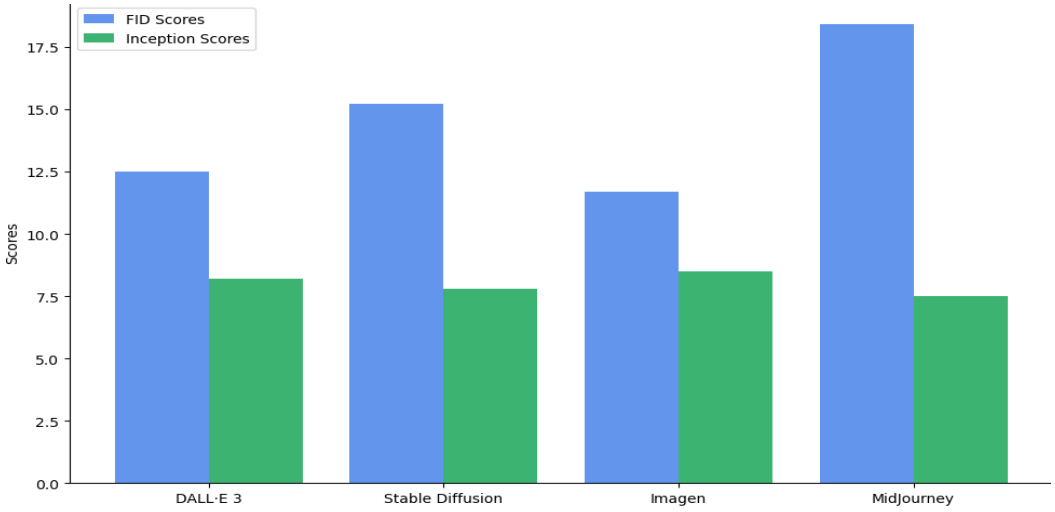

Figure 5. FID and Inception score [25].

4.3 Model Training and Inference

Each model was considered in its pre-trained state for performance out of the box. Additionally, we also looked at the possibility of fine-tuning further on domain-specific datasets.

- Stable Diffusion: This uses latent diffusion, which works by compressing data into a latent space, hence being more resource-efficient. Stable Diffusion was trained on LAION-5B and has become notable for its ability to generate highly detailed images without requiring significant hardware.
- DALL·E 3: This is an improved version of GPT-4 architecture in terms of language understanding and generating images. Although the dataset used for this model by OpenAI is proprietary, making replication impossible, it is pretty capable of generating creative, unique, and coherent images.
- Imagen: Imagen by Google is trained on the JFT-300M using a diffusion model. Imagen maintains a strong focus on photorealism to generate real-world scenes complete with minute details and elevated levels of visual accuracy.
- MidJourney: Fine-tuned and updated regularly for more creative and abstract visualizations. However, it primarily focuses on Artistic output since it is mostly oriented to a full artistic expression rather than sticking to the real-life object, it is extremely popular among imaginative designs.
- GANs: Generative Adversarial Networks such as BigGAN and StyleGAN involve training two neural network generators and a discriminator in an adversarial fashion [10]. While

GANs can yield highly realistic images, they often must be thoroughly trained and do not lend themselves to precision in text-to-image alignment compared with diffusion models. Text-to-GANs, or to be more precise, text-conditioned GANs, extend GANs for text prompts but usually struggle when the text input is complex. Figure 6 shows the Text-image alignment vs. Inference speed.

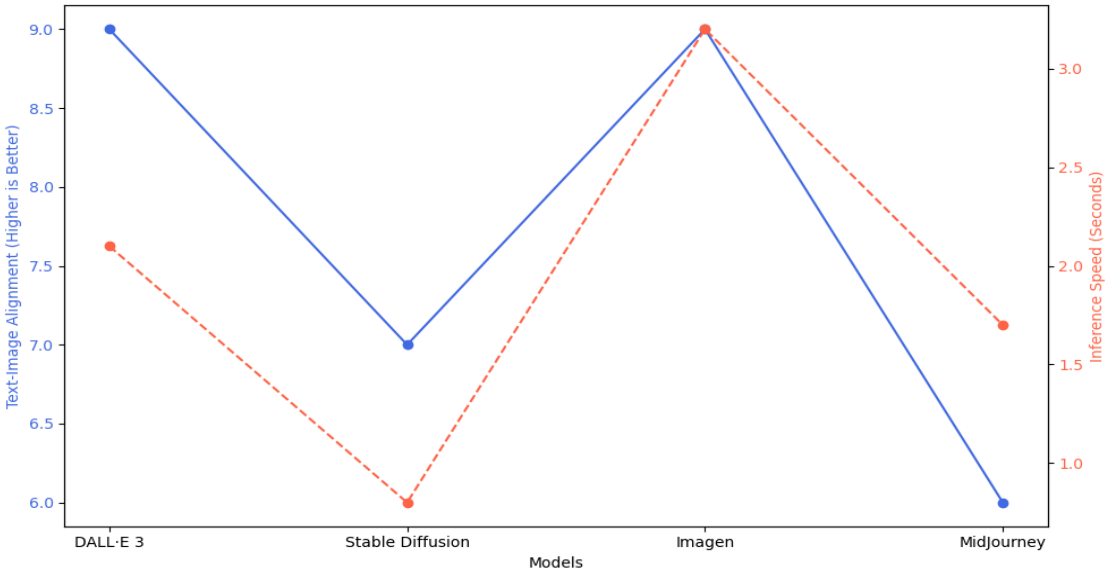

Figure 6. Text-image alignment vs. Inference speed [25].

4.4 Qualitative Evaluation

Qualitative test images were generated on all models from various text prompts. This has been done to experiment with whether each model can generate different scenes, objects, and complex and abstract ideas. Some examples include:

- Simple Object Descriptions: "A red apple on a white table."
- Complex Scenes: "A bustling street market at sunset, with people walking and vendors selling fresh produce."
- Abstract Concepts: "A surreal landscape with floating islands and cascading waterfalls."
- Text in Images: "A billboard in a city with the words 'Innovation for the Future' written on it."

The performance of GANs during this test was factored in, though one would note that at the time they were behind regarding the alignment of text and image compared to the results from diffusion models. But when it comes to standalone images, GANs are still extremely competitive about the quality of the visual output [7].

4.5 Architectural Comparisons

Architectural differences between the models formed another especially important aspect of our analysis. Indeed, each model represents a different approach to solving the task of text-to-image generation:

- **DALL·E 3**: combines transformers with **GPT-4**, a very sophisticated language model which, if used properly, allows gaining deep insights into complex prompts. The image decoder generates the final visual output, and here, the model shines in creative and imaginative tasks.

- **Stable Diffusion**: A latent diffusion model that works by progressive updates of the images through iterative steps of denoising in the latent space, thus enabling fast and resource-efficient image generation.
- **Imagen**: Comprises both diffusion models and attention mechanisms, doing quite well in capturing real-world details and hence achieving high-profile progressively by denoising.
- **MidJourney**: From an artistic, abstract generation proprietary model that targets it, it does poorly in photorealism but also has bad text alignment and is amazingly popular in creative results.
- GANs: The GAN architecture, inspired by models like BigGAN and StyleGAN, consists of a generator that creates images and a discriminator evaluating them with adversarial training to produce more realistic outputs. Text-conditioned GANs, such as AttnGAN, incorporate additional text embedding layers into the generation process to incorporate textual input but provide less precision than diffusion-based models [20]. Table 5 shows the Model comparison based on architecture, Datasets, and computational requirements.

Table 5. Model comparison based on architecture, Datasets, and computational requirements.

Model	Architecture	Dataset Used	Parameters	Training Resources	Inference Speed	Open-Source Availability
DALL·E 3	Transformer + GPT-4	Proprietary Dataset	12B	High (Proprietary)	Moderate	No
Stable Diffusion	Latent Diffusion Model	LAION-5B	890M	Moderate (Public)	Fast	Yes
Imagen	Diffusion + Attention	JFT-300M	1.5B	High (Proprietary)	Slow	No
MidJourney	Proprietary	Proprietary Dataset	Unknown	Moderate (Proprietary)	Moderate	No

4.6 Limitations

Although the given study attempts to provide an overall comparison, there is an intrinsic limitation within the approach:

- Subjective Bias: The quality and creativity in generated images can be subjective to human evaluation.
- Computational Constraints: Some models are not open-sourced-for instance, Imagen and DALL·E 3 which limits their customization and independent evaluation by researchers.
- Ethical Considerations: These are beyond this technical review, but ethical issues including bias in training data that AI models learn from are recognized as of paramount importance.

5. Challenges and Future Directions

5.1 Current Challenges in Text-to-Image Generation

Though a lot of development is being made in the field of text-to-image generation, at the larger scale of these models for general broad use cases, several challenges persist. Key challenges include:

5.1.1 Model Bias and Ethical Concerns:

Text-to-image models, among which DALL·E 3 and Stable Diffusion have their places, are typically trained on big, unfiltered datasets from the internet. Thus, such models have inherent biases

in their outputs. These biases may come in the form of stereotypes, cultural insensitivity, and overrepresentation of groups and ideas.

Ethical implications include potential misuse for generating deepfakes or inappropriate content, raising concerns about the responsible use of generative AI.

5.1.2 Image Coherence in Complex Scenes:

This is while models such as Imagen and DALL·E 3 are particularly good in photorealism, generating coherent images for overly complex or abstract text prompts remains a challenge. This includes the production of images that require a deep understanding of spatial relationships, object interaction, and multi-step reasoning.

5.1.3 Resource Intensity:

Most of the reviewed models, like Imagen and DALL·E 3, require high computational resource needs. As such, they become incompetent for deployment in a real-time application, let alone large-scale. The demand for high resources also translates to limited access because small research labs or startups cannot afford sufficient computational power to put these models into effective use.

5.1.4 Limited Open-Source Access:

Although Stable Diffusion and DeepFloyd IF are open source, most advanced models, such as DALL·E 3 and Imagen, are proprietary; this cuts off the wider research community from making valuable improvements. Open-source lack also slows down the development of tuned or domain-specific models.

5.2 Future Directions and Potential Improvements

The area of text-to-image generation has tremendous scope for innovation, and these lacunae are likely to be overcome by future developments.

5.2.1 Model Fine-Tuning and Customization

This, however, will further expand the scope of text-to-image generation with broader capabilities for fine-tuning the model on domain-specific applications such as medical imaging and industrial design. Future models could be fitted with easier fine-tuning mechanisms that enable non-professionals to adjust model parameters for unique tasks.

5.2.2 Ethical AI and Bias Mitigation:

In this respect, there is an urgent need for frameworks that take active measures against biases in generative models. The development can be done either by preparing filtered datasets for training or integrating bias detection algorithms into the model. That can be done only through deeper research in ethically aligned AI.

5.2.3 Improving Image Coherence and Compositionality:

Advances in multi-modal learning and scene understanding are foreseen to enhance the models' capability for richer interpretations and generations of complex scenes. Further, future models may involve 3D understanding or scene graph analysis that will provide better spatial relationships and dynamic interactions within the generated images [15].

5.2.4 Enhancing Efficiency

Stable Diffusion models already try to advance the state-of-the-art in computational efficiency, but there is still a lot more work to be done with the resource-efficient model development. This involves quantization techniques and lighter-weight architectures that can maintain high image quality by reducing the computational cost.

5.2.5 Exploring Real-World Applications:

With increased efficiency and ease of use, wider applications are envisioned in entertainment, advertising, education, and healthcare. Further research is encouraged toward domain-specific implementations that ensure safe and effective text-to-image generation integration in those fields.

5.3 Future Potential

While progress has indeed been phenomenal in text-to-image generation, the leading platforms remain DALL·E 3, Stable Diffusion, and Imagen. However, there are many challenges yet to be overcome, especially with respect to bias mitigation and ethical usage of such models. Moreover, scaling is one of the significant issues; hence, the prospect of future scaling models looks especially important. Further optimization will occur on fine-tuning, open-source collaboration, and computational efficiency, turning the tide toward a new era of creativity in AI [12]. Table 6 shows the Future directions and potential improvements. Table 7 shows Integration with other systems. Table 8 shows the Hardware requirements.

Table 6. Future directions and potential improvements.

Model	Key Improvement Areas	Potential Future Research Areas
Stable Diffusion	Bias mitigation, improved text alignment	Customization for niche applications, like healthcare or design
DALL·E 3	Reducing computational resource demands	Ethical AI frameworks, handling ambiguous prompts
MidJourney	Better handling of detailed text prompts	Improved scene generation and textual understanding
Imagen	Open-source accessibility	Increasing computational efficiency and scaling
DeepFloyd IF	Image coherence, more detailed prompts	Exploring artistic use cases and creative tools
eDiff-I	Resource optimization, fine-tuning	Better scene understanding, dual diffusion improvements

Table 7. Integration with other systems.

Model	Integration Capability	Supported Tools/APIs	Ease of Integration	**Comments**
Stable Diffusion	High	Yes	Easy	Open-source, runs on consumer-grade hardware, fine-tunable
DALL·E 3	Moderate	No	Moderate	Accessible only via OpenAI's proprietary API, limiting flexibility; requires internet access for API calls.
Imagen	Low	Google Cloud APIs (potential for integration)	Difficult	Closed source and limited to Google's ecosystem; no direct local deployment options.
MidJourney	Moderate	Discord Bot Integration	Easy	Requires integration via Discord; limited control over backend processes, but intuitive for non-developers.
DeepFloyd IF	High	Diffusers (Hugging Face), PyTorch	Easy	Open-source and designed for research and experimental use; supports integration with popular deep learning libraries.
GAN Models	Moderate to High	TensorFlow, PyTorch, custom pipelines	Moderate to Difficult	Requires significant customization for text-to-image tasks; flexible but demands expertise for effective integration.

Table 8. Hardware requirements.

Model	Training Hardware Requirements	Inference Hardware Requirements	Comments
Stable Diffusion	Moderate: 8+ GB VRAM, NVIDIA RTX 2080, or equivalent, large storage for datasets (e.g., LAION-5B)	Low: 4+ GB VRAM, runs on consumer-grade GPUs like GTX 1060 or better	Highly efficient, designed to run on consumer hardware; suitable for local deployment.
DALL·E 3	High: Proprietary large-scale cloud infrastructure (multi-GPU/TPU clusters)	Moderate: Requires access to OpenAI servers; specific GPU requirements unknown.	Cloud-dependent; not accessible for local or offline usage.
Imagen	High: Google TPU Pods, proprietary JFT-300M dataset storage	High: Google cloud-based infrastructure	Only accessible via Google's proprietary systems; not open-source or available for local use.
MidJourney	Moderate: Proprietary training environment, GPU clusters	Moderate: Accessible via cloud-based subscription service	No public details on hardware specifics; inference is handled entirely on Midjourney's servers.
GAN Models	Very High: 16+ GB VRAM, multi-GPU setups, high storage for datasets	High: 8+ GB VRAM for advanced GANs like StyleGAN	Requires careful optimization; typically, resource-intensive both for training and inference.

6. Results and Discussion

This would be the review section of the performance of each of the text-to-image generation models, concerning the evaluation metrics and graphs discussed above. The problems involved in this comparison would lie along multiple axes, such as image quality, computational efficiency, and text-image alignment.

6.1 Image Quality (FID Score and Inception Score)

From the bar chart, the FID score and Inception score yield insight into the overall quality and diversity of images produced using the models.

- Imagen does have the best FID score at 11.7, meaning that this generates the most realistic images, which is expected as it makes use of diffusion models combined with attention mechanisms optimized for photorealism.
- DALL·E 3 will also impress with its FID score of 12.5 and an impressive Inception score of 8.2, showing a strong balance between the diversity of images and their quality.
- Stable Diffusion is similarly respectable with an FID score of 15.2 and an Inception score of 7.8, though a bit worse in terms of realism compared to Imagen and DALL·E 3.
- MidJourney, while creating a higher output in terms of creativity and artistry, had the highest FID score at 18.4, showing its images are less real but varied.

6.2 Text-Image Alignment and Inference Speed

The line graph of Text-Image Alignment versus Inference Speed shows the trade-offs of the models.

- DALL·E 3 and Imagen attain close-to-perfect text-image alignment, with scores of 9/10, which is indicative of good comprehension by models and the generation of images that match the provided text prompts. However, both have much slower inference, especially Imagen, at 3.2 seconds on average, hence much more computationally intensive.
- Stable Diffusion yields a little lower alignment score, 7/10, but this model exhibits extremely fast inference speed, averaging 0.8 seconds. Thus, it can be utilized in cases where the speed and efficiency of the diffusion process are crucially needed, such as real-time applications.
- With text-to-image alignment, MidJourney performs poorly, with only 6/10. Again, it provides proof of the assumption that it is more creative yet less accurate. The inference time of MidJourney was taken from the middle, compared to other models, with an average of 1.7 seconds.

6.3 Model Accessibility and Scalability

Among other key features are the accessibility and scalability of each model:

- Stable Diffusion, on one side, is open source; hence, many developers and researchers from different circles can use and further develop this model. The scalability and computational efficiency of Stable Diffusion make it very versatile in many applications-from creating art to content generation and even medical imaging.
- While both DALL·E 3 and Imagen gave superior performance in terms of alignment and image quality, they remain proprietary and resource-intensive. For this reason, most members of the public cannot have access to this technology, nor are large-scale experiments feasible with these models.
- MidJourney is proprietary but accessible through subscription, so the artist and designer community has taken quite a liking to it.

6.4 Applications and Use Cases

Each of these models excels in various applications, including the following:

- Stable Diffusion: Best fitted for developers who want flexibility, customizability, and speed for real-time applications like mobile apps and other interactive platforms.
- DALL·E 3: Fitted for creative professionals who require fine image quality, contextual accuracy, and an elevated level in the areas of marketing, storytelling, and branding.
- Imagen: Ideal for photorealistic image generation in e-commerce, product design, or scientific visualizations.

- MidJourney is more in line with artists and designers who feature the creative aspect of and are artistic in their approach rather than photorealism.

7. Conclusion

In this paper, we explored the fast-moving domain of text-to-image generation, focusing on Stable Diffusion, DALL·E 3, Imagen, and MidJourney. We investigated in detail the architecture, datasets, and performance metrics of every single model. That will help us to show the exclusive strengths and limitations typical for each model. This comparative study becomes necessary to identify an application-specific model that will satisfy the specific needs that arise either from research, industry, or creative needs.

Stable Diffusion is an unusually accessible and efficient model; it is extremely versatile for a vast variety of use cases. It is open source; hence one can modify it and toy with it. Its computational requirements are low, which is important to scale real-world usage. On the other hand, its open-source counterpart presents a slight trade-off compared to proprietary models such as DALL·E 3 and Imagen, with slight trade-offs in image quality and text-image alignment. This makes it an excellent fit for many practical applications despite these differences, featuring its great speed and adaptability, especially in real-time or resource-limited environments.

On the other hand, DALL·E 3 and Imagen push the boundaries of image quality and text-image fidelity; hence, excellent in photorealistic generation and interpretation of complex prompts. These models show the state of the art, especially in professional settings where image quality and context accuracy matter a lot. However, larger computational cost, the use of proprietary datasets, and limited accessibility restrict their potential wide diffusion on a large scale. Specialized applications best suit these in industries like marketing, e-commerce, and scientific visualization where great realism and detail are essential.

MidJourney, even though its focus was not on photorealism, carved a niche for itself in creative industries. Providing highly artistic and abstract content, it has emerged as a favorite tool among designers and artists. While its scores on text-image alignment and realism are low in comparison to DALL·E 3 and Imagen, the capability for derivation of an exceptional view along with stylistic diversity makes the outputs unique in fields requiring subjective and more imaginary content generation.

In this view, with further developments of AI-based text-to-image models in the future, much emphasis should be put into enhancing model explainability and ethical issues while reducing computational obstacles in accessing state-of-the-art models like DALL·E 3 and Imagen. This also opens opportunities for models like Stable Diffusion to be fine-tuned on domain-specific applications, including healthcare, education, and entertainment, where demands for high speed and quality are indispensable.

This will boil down to the needs of the application speed and accessibility are more important with Stable Diffusion or when superior image fidelity and alignment via DALL·E 3 and Imagen are key. As text-to-image generation technology grows more mature, it will be an indispensable tool across industries and a shaper of the future of content creation in this visual world.

Declarations

Ethics Approval and Consent to Participate

The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration by another publisher. All the material is owned by the authors, and/or no permissions are required.

Consent for Publication

This article does not contain any studies with human participants or animals performed by any of the authors.

Availability of Data and Materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing Interests

The authors declare no competing interests in the research.

Funding

This research was not supported by any funding agency or institute.

Author Contribution

All authors contributed equally to this research.

Acknowledgment

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

References

[1] A. Ramesh, M. Pavlov, G. Goh, et al., "DALL E: Creating Images from Text," arXiv preprint arXiv:2102.12092, 2021.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv preprint arXiv:2112.10752, 2022.

[3] C. Saharia, W. Chan, S. Saxena, et al., "Imagen: Text-to-Image Diffusion Models with Large Pretrained Language Models," arXiv preprint arXiv:2205.11487, 2022.

[4] A. Nichol, P. Dhariwal, A. Ramesh, et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," arXiv preprint arXiv:2112.10741, 2022.

[5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv preprint arXiv:1312.6114, 2013.

[6] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," arXiv preprint arXiv:1503.03585, 2015.

[7] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models from Natural Language Supervision," in Proc. International Conference on Machine Learning (ICML), 2021.

[8] Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," arXiv preprint arXiv:2006.11239, 2020.

[10] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[11] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv preprint arXiv:1411.1784, 2014.

[12] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980, 2014.

[13] T. Salimans, I. Goodfellow, W. Zaremba, et al., "Improved Techniques for Training GANs," arXiv preprint arXiv:1606.03498, 2016.

[14] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," arXiv preprint arXiv:1809.11096, 2018.

[15] X. Ren, Z. Lin, Z. Lu, et al., "Deep Generative Models of 3D Shapes with Random Fields," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[17] DeepFloyd AI, "DeepFloyd IF: Text-to-Image Generation with Transformer and Diffusion Models," arXiv preprint arXiv:2301.08918, Jan. 2023.

[18] C. Schuhmann, R. Beaumont, and R. Vencu, "LAION-400M: Open Dataset for Large-Scale Multi-Modal Research," arXiv preprint arXiv:2111.02114, Nov. 2021.

[19] NVIDIA Research, "eDiff-I: Enhanced Diffusion Models for Image Synthesis," arXiv preprint arXiv:2301.08074, Jan. 2023.
[20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in Proc. Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 6626-6637.
[21] https://medium.com/@alextakele16/image-generation-using-stable-diffusion-mo-b3cf65481692
[22] https://imagen.research.google/paper.pdf
[23] https://medium.com/@tusharpawar749963/generate-what-you-imagine-unleashing-the-power-of-stable-diffusion-and-comfy-ui-c0fbc63a1465
[24] https://www.researchgate.net/figure/The-structure-of-Latent-Diffusion-Models-30_fig2_382384724
[25] https://github.com/Abdallah-007/Image-Generation

Received: 01 Oct 2024, **Revised:** 28 Feb 2025,
Accepted: 24 Mar 2025, **Available online:** 26 Mar 2025.

© 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.