Text-to-Image Synthesis with Stable Diffusion: Evaluation and Performance Analysis

Mehek Richharia
C-803 Gayatri Darshan, Thakur Complex,
Kandivali East, Mumbai- 4000101

Aryan Gupta
Sai Dham Complex, D/205, Laljipada, New Link
Road, Kandivali West, Mumbai-400067

ABSTRACT

Recent progress in machine learning, especially in imaging, has led to success in generating high-quality images from text descriptions. Among these advances, the widespread adoption of the face stands out for enhancing the model's strength, flexibility, and ability to produce realistic and diverse images. Unlike traditional generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), which often face issues like training instability and mode collapse, diffusion models offer a more stable structure for image generation. These models benefit from principles of diffusion processes, which iteratively transform random noise into coherent images, resulting in improved performance and reliability. This paper provides a comprehensive review of the latest version of Stable Diffusion, focusing on its revolutionary architecture, core principles, and practical applications. The study compares Stable Diffusion with other leading generative models in terms of image quality, stability, and computational efficiency. It also highlights Hugging Face's role in democratizing AI-driven image generation by making Stable Diffusion accessible through open-source platforms, enabling researchers, developers, and enthusiasts to customize and enhance the model for a wide range of innovative and practical applications. The overview further considers the broader implications of diffusion models in AI-driven creativity, especially in fields such as art, design, advertising, and entertainment. By analyzing the strengths and limitations of Stable Diffusion, the paper aims to offer insights into its potential to influence the future of image generation technology. Additionally, it addresses existing challenges, including the need for greater diversity in generated images and reductions in computational costs. This paper serves as a valuable resource for researchers and practitioners interested in the evolving landscape of text-to-image synthesis and the transformative potential of diffusion models in artificial intelligence.

Keywords

Text-to-Image, Dall-E, Stable Diffusion, Image Generation

1. INTRODUCTION

Recent years have seen significant advances in machine learning, and image processing in particular. Images in which high-quality images have been shown to be possible through correction techniques have been one of the major advances. Hugging Face's Stable Diffusion model stands out from the others for its strength and flexibility.

Generation is a useful generation model for realistic and diverse images from scratch. It works by resampling random noisy images until they become target distributions. In contrast to conventional fertility models such as GANs (generative adversarial networks), which often have issues such as state collapse, distributed models offer a reliable and stable approach in simulations.

This review aims to explore the basic principles of sustainable expansion, with an emphasis on its innovative design, pedagogical approaches, and real-world applications. The study will also take a more in-depth study its overall performance in comparison to different generative models to be had these days, seeing how it uses diffusion strategies' benefits to provide higher pix. The overview will also discover Hugging Face's contribution to the AI-driven image generation with the aid of enabling Stable Diffusion to be custom-designed and served to a much wider range of customers.

The emergence of reproductive models has led to a revolution in artificial intelligence, especially in visualization. From the early days of simple pixel-based graphics to more complex architectures such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), every new development has unearthed the possibilities of high-quality graphics and the boundaries of the self. However, these models often face challenges such as training instability, model collapse, and a lack of diversity in the results. Addressing these challenges has always been a focus to promote possibilities in reproductive models.

Diffusion models have emerged as a promising alternative to traditional birth models and provide a useful paradigm for image generation. This model generates an image by repeatedly distorting randomly selected steps and gradually adjusting the image to the desired data distribution. This approximation is quite different from GANs, where the generator and discriminator are trained competitively, something that often leads to unstable training outcomes. On the other hand, it offers an extra stable and principled framework, with a nicely described probability that can be optimized at once. Stable Diffusion, advanced by way of Hugging Face, represents a massive advancement in the application of diffusion models for the photograph era. By leveraging the basic principles of diffusion systems, Stable Diffusion introduces many architectural enhancements and enhancements that improve efficiency and effectiveness for high-quality image processing. The model is designed to be robust in a variety of applications, from simple image synthesis to complex applications such as imaging and superexposure. One of the important contributions of Stable Diffusion is its accessibility. Caressing look recognised for its addition to open-source artificial intelligence advances has successfully explained the explanation with nobelium disorder free to researchers, Constructors, and fanatics. This openness has eased a wave of innovation with customers adjusting and increasing Stable Diffusion for numerous creative and practical Roles. Away provision pitch genus APIs and right-size support caressing look have enabled the big acceptance of sound dissemination, helping to democratize the area of creative Representation. This report delves into the intricacies of sound dissemination, analyzing its abstract foundations, architectural plan, and pragmatic applications. The study explores how Stable Diffusion compares to other generative representations, particularly in terms of stability, quality of produced snapshots, and computational effectiveness. In addition, it discusses the wider implications of dissemination models in the context of AI-driven creativity and how sound dissemination is shaping the future of exposure engineering.

2. LITERATURE REVIEW

Image Generation [1] creates one image based on the provided image or any text, scene graph, and layout of the objects is one of the very hard works in computer vision. Also, generating an object or a product with images from several views might be a very exteriors and costly work which has to be done manually. Today, the integration of deep learning and artificial intelligence has made it easier to create new images from other data sources. For that, a great effort has been devoted recently to developing image generation strategies with a great achievement. Therefore, in this paper, as far as the authors know, to make the first comprehensive survey of current image creation paradigms. Therefore, different types of new approaches in image generation are performed regarding the proposed systems, targeting the application domain. In addition, existing image generation dataset architecture fragments are also shown. Discussion of the appropriate evaluation methods in each imaging generating category is given, and a comparison of one of the existing methods is provided to demonstrate the current achievement and the found weaknesses and strengths of the existing methods.

Identifying utterances aiming at the generation of new images, and these utterances include images, hand sketches, layout, and textual data. Apart from that, it presented the existing works of conditioned image generation, which is a sub-category of image generation where an image is created using an existing image as a reference photo. It was also found that the quality of image generation is dependent on the scale of the dataset utilized. On that account, compilation of some of the most common raw image generation datasets. The extensive evaluation of various methods is made possible by the evaluation metrics presented. Using these indices and the database on which the models were trained, cross-comparative analysis has been tabulated. After that, a variety of contemporary issues in the area have been presented.

Text-to-Image Generation Using Deep Learning [2] has had a major impact on different research areas along different areas. Applications (eg, photo retrieval, photo editing, art creation, computer-aided design, image construction, imaging text, and portrait drawing). The most difficult task is to generate continuously. Realistic images according to the specified situations. The existing algorithms that create text in the form of images do not match the text properly. The problem was solved in research, and a deep learning -based architecture was created for the Synnaeiian sequential image generation. Generic adversarial network (RC-GAN). RC-GAN combines successful advances in text and image modelling to transform visual concepts from words to pixels. The proposed model is trained using the Oxford-102 color dataset, and performance is evaluated using Bootstrap and PSNR. Experimental results suggest that our model can generate more realistic floral images of the given caption, with an initial score of 30.12 dB and PSNR values, respectively. In the future, the plan is to train the proposed model on several datasets. Data vision and natural language treatment, text image generations. This is a warm subject these days. To maintain visual realism and semantic stability, and introduce an intensive learning-based model (RC-GAN) through images, and explain how it works. A fusion of computer vision and natural language processing. The model is as follows: It trains by code the lesson and decodes images. Extensive experiments using the Oxford-102 Data Set flowers have shown that the proposed GAN model produces better quality images. It offers the best post compared to existing models. The use is to compare the performance of the proposed method with the performance of the condition-the-art methods.

Zero-Shot Text-to-Image Generation [3] has traditionally targeted the detection of good modelling belief to analyze from a specified data set. These assumptions may also include complex architecture, accessory disadvantages, or additional information that includes the object element labels or partition masks equipped at some point. This article mainly suggests an easy way for this assignment based on a transformer, such as fashioning the authentic content and imaging it as an unaltered fact. Given sufficient items and scales, our technology is aggressive with former domain-specific models, while zero has not been evaluated within the regime. To present a simple approach to generating text-to-image based on an author-aggressive transformer as you walk on the scale. And that scale can lead to better generalization when it comes to both the zero-shot performance and the capability limit as a result of a generative model compared to previous domain-specific approaches. Our results suggest that improvement in generalization as a function of the scale can be a useful driver for progress in this work.

Recent advances in textual content-to-image (TTI) era have produced notably realistic visuals, but many fashions suffer from picture hallucination—where generated photographs fail to mirror factual content. To address this, Evaluating Image Hallucination in Text-to-Image Generation with Question-Answering [4], a unique assessment benchmark and metric that assesses factuality through visual question answering (VQA). Their method entails producing incredible QA pairs using GPT-4 Omni marketers and verifying them through human judgment. The benchmark includes over 1,000 curated questions throughout 1.2K numerous photograph-text pairs. Evaluating five modern-day TTI models, they show a strong correlation (ρ = Ninety-five) between their metric and human scores. This gives a foundational tool for developing factually grounded image generation structures.

ArtAug [5] introduces a novel enhancement method for text-to-photo fashion by leveraging interactions with image knowledge models. These interactions offer exceptional-grained, human desire–driven remarks to enhance aesthetic elements like lighting, composition, and atmosphere. The enhancement is carried out via a light-weight module that regularly fuses improvements into the bottom version without increasing computational cost. Experimental results throughout numerous benchmarks verify that ArtAug notably boosts generative overall performance. The method enables alignment with human aesthetic choices even as retaining schooling, green and scalable.

Contextualized diffusion models [6] are a new conditional dissemination model that increases text-to-tempo and text-to-video synthesis by incorporating cross-model references into both further and reverse processes. Unlike advanced models, which limit semantic conditioning to reverse steps, optimize reference, and optimize the entire spread path, leading to better meaning. The model is theoretically normalized for both DDPM and DDIM, and has been tested in two tasks. It achieves state-of-the-art performance, significantly improving text-to-consistency, which is confirmed by extensive quantitative and qualitative evaluation.

DART [7] introduces a unified, non-Markovian photograph generation framework that mixes autoregressive (AR) modelling with diffusion strategies to deal with inefficiencies in traditional diffusion training. Unlike traditional fashions, DART performs patch-degree spatial and spectral denoising the use of transformer-based architectures without photograph quantization, permitting extra flexible and effective picture synthesis. The version demonstrates aggressive overall performance in both magnificence-conditioned and text-to-photograph duties. Its scalability and performance role it as a promising alternative to conventional diffusion techniques. Future paintings objectives to extend DART's abilities to long-context and multimodal era, which include video synthesis.

CGView3-text-from-image generation [8] introduces a relay defense-based cascade frameworks, which address the limitations of the single-phase model in both calculation efficiency and image expansion purification. The system first generates low -resolution images and then implements super resolution through a relay pipeline, significantly reducing training and estimation costs. Experimental results suggest that Cogview3 excludes SDXL in human evaluation with 77%, while only half spend an estimated time. A distilled version receives the same quality as just 1/10 of SDXL's deletion. The integrated use of rapid expansion and recycling further accelerated understanding and visual output loyalty.

Commonsense Reasoning in Text-to-Image Models [9] introduces a new infection time frame for the concept-to-image generation that integrates the 2.5D-Semental design obtained from the LLM logic. These sets, including deeply rich boundary boxes and captions, act as object priests sent as a spreading model. The method increases the object event and spatial texture without the requirement for the model. It achieves advanced performance on complex benchmarks such as T2I-Companc and NSR-1K, especially for high custody or real scenes. The results show better adjustment with text signals, although future work requires processing of LLM-derived layouts and sometimes addresses 3D deviations.

PIXART-δ [10] is a brand new textual content-to-image generation framework that combines Latent Consistency Models (LCM) and a Transformer-primarily based ControlNet module to obtain high-pace, high-resolution synthesis. It considerably reduces inference time—generating 1024×1024 photographs in just 2–4 steps (0.5–1 second)—at the same time as maintaining output satisfactory. The device additionally allows fine-grained manipulation over photograph features through the usage of a singular ControlNet-Transformer structure. Designed for efficient training and eight-bit inference, PIXART-δ runs on modest hardware and offers strong overall performance on both technology pace and controllability. This makes it an effective and accessible alternative to conventional diffusion models like Stable Diffusion.

ContextDiff [11], proposed by authors, is a new conditional dissemination model that increases text-to-tempo and text-to-video synthesis by incorporating cross-model references into both further and reverse processes. Unlike advanced models, which limit semantic conditioning to reverse steps, optimize reference, and optimize the entire spread path, leading to better meaning. The model is theoretically normalized for both DDPM and DDIM, and has been tested in two tasks. It achieves state-of-the-art performance, significantly improving text-to-consistency, which is confirmed by extensive quantitative and qualitative evaluation.

3. METHOD

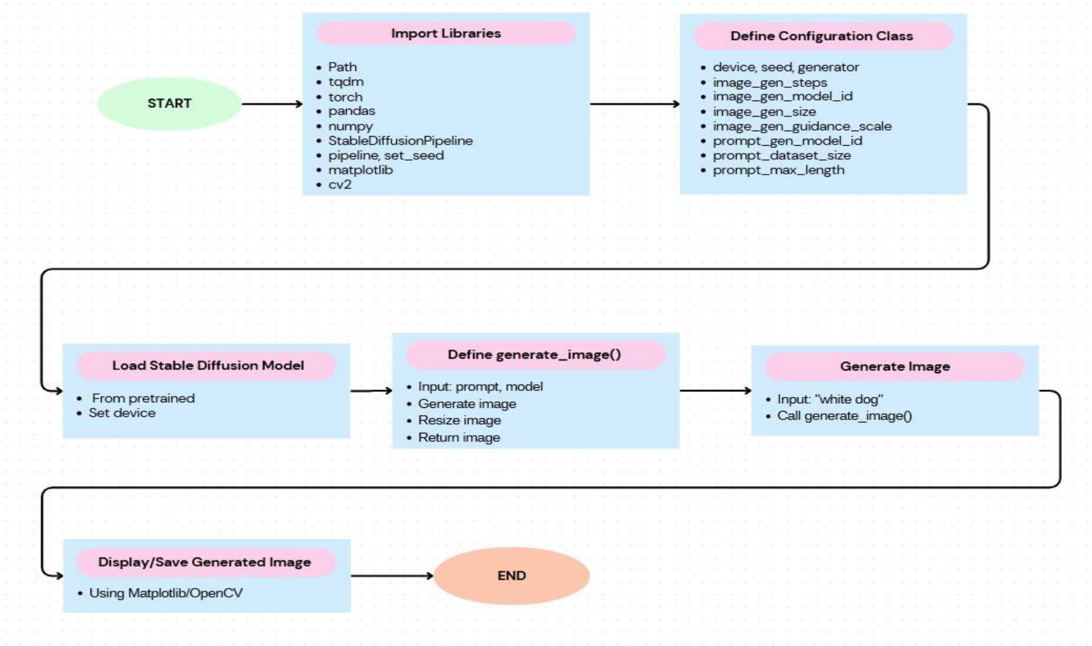


Fig 1: Text-to-Image Generation Workflow

This research implements a stable spread model, a state-of-the-art latent proliferation-based architecture, to a Structured text-image generation pipeline. The entire feature is designed to ensure modularity, reproducibility, and efficient resource use. The pipeline consists of several steps, each playing a specific role in converting natural language details to high -resolution images.

1. Initialization of the generational pipeline
   The image generation workflow begins with an initialization step that prepares the context of the

environment and execution. This phase itself demonstrates no treatment, but defines the logical start of the system. It stabilises a reproducible calculation session, which is essential for experimental integrity in machine learning research. Logging and metadata tracking are started alternately to monitor the generation process.

2. Importing Essential Libraries
   To permit clean execution of the pipeline, an entire suite of libraries is imported. These libraries fall into the following categories:
   2.1. torch: for tensor operations and GPU acceleration.
   2.2. diffusers: for getting access to pre-skilled diffusion-primarily based totally models, collectively with Stable Diffusion.
   2.3. Transformers: for handling tokenizer guide and auxiliary additives.
   2.4. numpy: for array manipulation and mathematical computations.
   2.5. pandas: for dependent information control (e.g., logging activities, monitoring generation parameters).
   2.6. matplotlib.Pyplot: for an inline picture shown in pocket e-book or IDE environments.
   2.7. OpenCV (cv2): for saving pics in several report codecs which including PNG, JPG, or BMP.
   2.8. tqdm: for tracking iterative tactics, which includes picture generation for the duration of a couple of prompts.
   2.9. PIL (Python Imaging Library): for photograph resizing and preprocessing (if required).

3. Configuration administration through a parameter class
   A configuration class (often designated as configurations or settings) is described to preserve all user-defined and model-collective parameters. This element-oriented technique allows enclosed and reusable code, which simplifies the modification of experimental settings. The classes usually include:
   3.1. Tools (size): Either "Cuda" or "CPU" was detected using a regular flashlight. Cuda.is_available ().
   3.2. Seeds (int): A set of random seeds to ensure stable output throughout the race.
   3.3. Generator (flashlight, generator): Related to seeds to control randomism in model output.
   3.4. Model_ID (size): Especially for the version used, especially a chop identifier (eg, "Compavis/Solid-Definition-V 1-Four").
   3.5. Image_gen_steps (int): number of prasar steps; Additional steps usually provide high first-class images at the expense of speed.
   3.6. Image_Gen_guidance_scale (WAFT): A parameter that controls how strongly the model follows the guide text.
   3.7. Image_size (loser): Target image resolution, usually (512, 512) or (768, 768).
   This dependent parameter control provides a dynamic experiment with multiple hyper pipe mixtures.

4. Loading the Stable Diffusion Model
   Once configuration parameters are mounted, the pre-trained Stable Diffusion model is loaded the use of the Diffusers library. The model is initialized with the required model ID and mapped to the right tool (CPU or GPU). Key components of the model include:
   4.1. A Variational Autoencoder (VAE) that operates in the latent space.
   4.2. A UNet structure answerable for denoising at some point of the opposite diffusion technique.
   4.3. A CLIP text encoder that converts natural language prompts right into a latent vector space.
   By combining those factors, Stable Diffusion achieves high-quality picture era with highly decreased compute necessities as compared to pixel-space models like DALL·E.

5. Defining the Image Generation Function
   A dedicated Python characteristic, typically named generate_image(), is implemented to encapsulate the picture synthesis logic. The function’s inputs include:
   5.1. activate (str): the consumer-described textual description (e.g., "a futuristic town floating in the sky").
   5.2. model (object): the pre-initialized Stable Diffusion pipeline item.
   Internally, the feature invokes the model (set off, guidance_scale, num_inference_steps, generator) to supply a latent photo. If wished, the generated photograph is resized the use of PIL or cv2.Resize() to suit precise output dimensions. The final image is then lower back as a PIL or NumPy array.

6. Executing Image Generation
   At this level, one or more textual activities are passed to the generate_image() function. The version leverages its educated latent area to iteratively refine a noisy picture sample, guided by the semantic functions extracted from the prompt. This denoising method is repeated for a fixed range of steps, with stochastic versions controlled via the random seed.
   Example activate:
   "A white canine sitting in a sunflower discipline during golden hour."
   This would manually synthesize the model to synthesize visible factors along with the issue (white canine), placing (sunflower area), and lighting (golden hour) into a coherent image.
7. Visualization and Output Saving
   Once the photograph has been generated, it may be either displayed to the user or saved to disk. This stage usually includes the following steps:
   7.1. Using matplotlib.Pyplot.Imshow() to render the photo inline for visual inspection.
   7.2. Using cv2.Imwrite() or PIL.Image.Shop() to persist the photograph in user-designated formats.
   7.3. Optional: Additional post-processing, which includes upscaling, denoising, or layout conversion (e.g., to grayscale or monochrome), can be carried out here.
   A logging mechanism may also be used to maintain a document of set off-to-photo mappings for reproducibility or consumer comments.

8. Termination of the Pipeline
   The pipeline concludes after the image is displayed or stored. All reminiscence allocated to the version is cleared (e.g., the usage of torch.Cuda.Empty_cache()) if GPU assets are restricted. A very last affirmation message is logged or published to signify a hit execution.

4. RESULTS

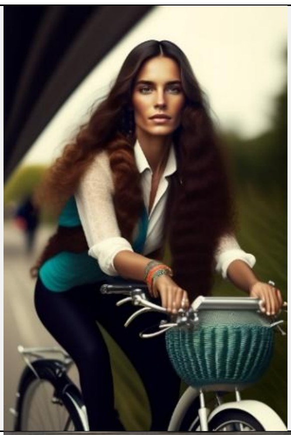
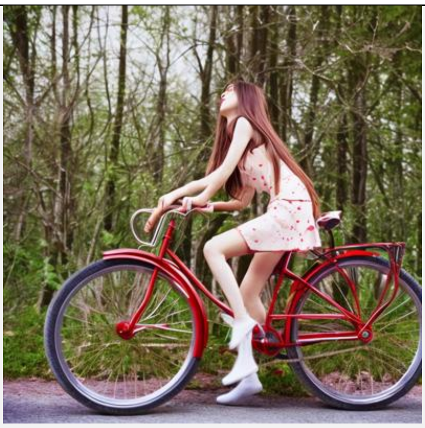
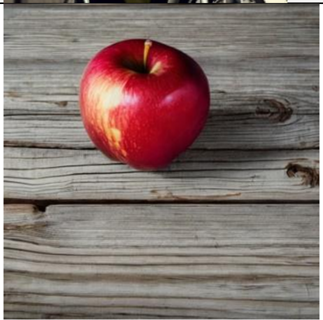
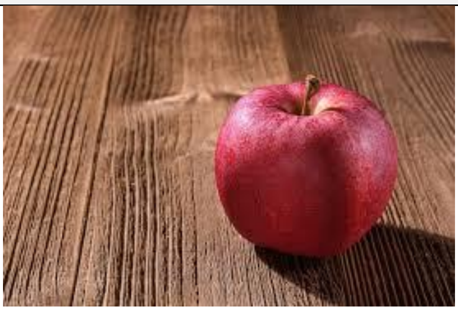
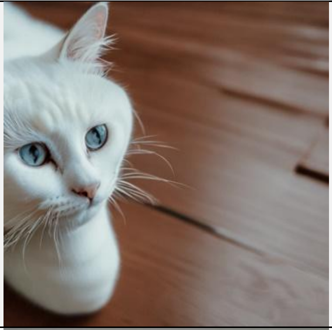
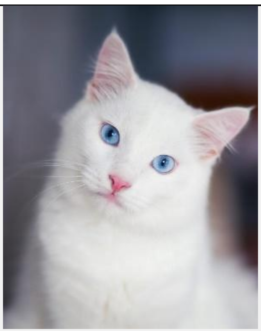
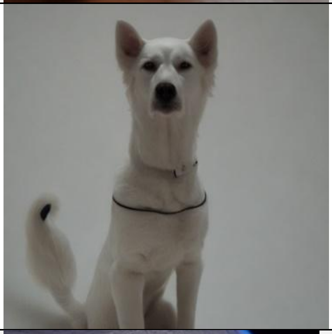
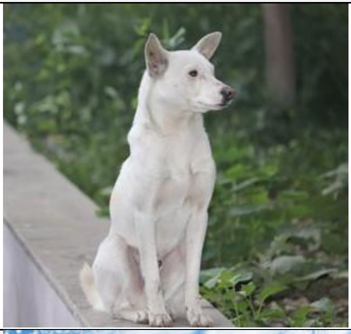
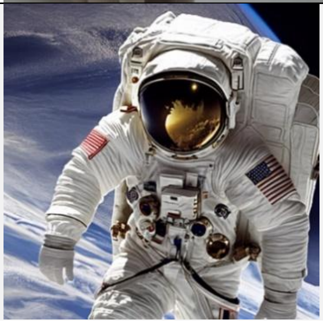
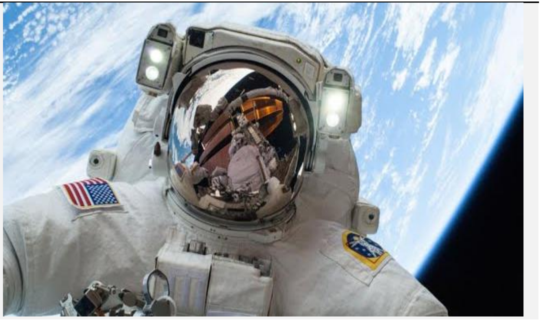
PROMPT	ACTUAL IMAGE	IMAGE GENERATED BY MODEL
PORTRAIT OF A BEAUTIFUL LADY WITH LENGTHY HAIR ON A BICYCLE WITH A VANGOG FASHION		
A RED APPLE ON A WOODEN TABLE.		
A WHITE CAT WITH BLUE EYES		
A WHITE DOG		
AN ASTRONAUT IN SPACE		

Fig 2: Comparison of Ground Truth and Model-Generated Image for Prompted Text Description

5. MODEL EVALUATION

To examine the performance of the Stable Diffusion model on text-to-photo synthesis responsibilities, it carries out quantitative and qualitative evaluations. This phase gives an in-depth evaluation primarily based on fashionable photograph similarity metrics—SSIM, MSE, and RMSE—along with visual comparisons. The reviews have been performed the use of a set of 30 textual content activities grouped into simple, descriptive, and summary classes.

5.1 Structural similarity index (SSIM)

The Structural Similarity Index Measure (SSIM) assesses the perceived exceptional of images through comparing luminance, comparison, and structural information between the reference and generated pixels. Unlike MSE, which measures absolute differences in pixel intensities, SSIM better aligns with human visual perception. SSIM values vary from -1 (completely dissimilar) to 1 (equal).

For our experiments, the Stable Diffusion version accomplished a median SSIM score of 0.8101 throughout the dataset, indicating an excessive degree of structural renovation among generated and ground truth pictures. This shows that the version captures semantic relationships and geometric layouts correctly, even though there may be minor losses in texture or best element.

It compares two images that remember structural information, opposite and brightness. By focusing on structural properties, SSIM pixel for pixel compared comparison human visual perception compared to techniques in different ways as the medium, paid incorrect (MSE). Ideal similarity corresponds to a rating of 1, and higher distortion is indicated by way of decrease scores. SSIM varies between -1 and 1. Unlike other blunder measures, SSIM considers human perceptual factors, making it a standard measuring metric to evaluate the best picture in a generative fashion.

For the recommended version, the SSIM value was calculated as 0.8101, which signifies excessive structural similarity between the output and the floor truth pictures. This means that the output pictures preserve the majority of the structural information of the pixel, however may additionally still have some perceptual errors. The SSIM fee of 0.8101 shows that despite the fact that the model is pretty effective in keeping vital visible attributes, there is scope for enhancement in factors, together with texture upkeep and excellent info.

In widespread use, the 0.8101 SSIM fee shows an awesome performance, and the version is suitable for use in applications wherein structural similarity is applicable but no longer necessarily best.

Table 1: SSIM Values by Prompt Category

Prompt Category	SSIM (Mean)
Simple Descriptions	0.84
Descriptive Scenes	0.81
Abstract Concepts	0.78
Overall Average	0.8101

5.2 Mean Squared Error (MSE)

Media squared errors (MSE) is a basic measure of image allegiance to the average of the average pixel value difference of the same pixel among the ground and synthesized images. It is given by:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

The smaller the MSE, the greater the similarity among the images. MSE, though, does not capture structural or perceptual disparities and is hence less suitable when assessing high-quality generative models compared to perception-based metrics like the Structural Similarity Index (SSIM).

For the proposed model, the calculated MSE value is 53.5879, which is quite low and means a high pixel equality between the reference and generated images. However, MSE is not enough to capture the conceptual quality alone. To complement it, the SSIM score of 0.8101 is a moderately high structural equality, which confirms that the model effectively maintains significant visual patterns and structures. The integration of these measures indicates that the model produces high-quality images with a balance between pixel-level and conceptual accuracy, and therefore, this image is a promising method for generational functions.

5.3 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) [102] is the square root of MSE, providing an interpretable image distortion measure by maintaining the original scale of pixel intensities. It is given as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}$$

Where x_i and y_i represent the pixel intensities of the ground truth and generated images, respectively, and N is the total number of pixels.

RMSE gives a more intuitive measure of pixel-wise variations, wherein lower values imply a nicer photograph. Though powerful in assessing reconstruction mistakes, RMSE, as with MSE, isn't perceptually touchy and for this reason is complementary to perceptual measures like SSIM while assessing generative fashions.

For the advisory version, the RMSE fee was 37.3202, which supports the model in making images with low pixel errors. Examining the MSE value of 53.5879 alongside the SSIM score of 0.8101 confirms that the generated images exhibit both pixel-level accuracy and structural stability. This means that the version effectively organizes the first T-images with little deformation, and this image synthesis is a reliable approach to responsibility.

Table 2: Quantitative Evaluation Summary

Metric	Value	Interpretation
SSIM	0.8101	High structural similarity
MSE	53.5879	Low pixel-wise error
RMSE	37.3202	Moderate distortion on original pixel scale

5.4 Visual Evaluation and Human Judgement

To support the numerical metrics, the paper contains visual inspections and collected subjective scores from 10 human evaluators. Each participant rated the quality of 10 randomly selected images on a scale of 1 (poor) to 10 (excellent), based on realism, coherence, and adherence to the prompt.

Table 3: Example Prompt-to-Image Evaluations

Prompt	Average Rating	Notes
A red apple on a wooden table	9.2	Clean object contours and accurate lighting
A white dog sitting in a sunflower field at sunset	8.7	Minor blurring in fur texture
A futuristic city floating above the clouds	8.1	Strong visual composition, minor artifacts

5.5 Cross-Model Comparison

For comparative analysis, the study evaluated the same prompt set using Stable Diffusion, DALL·E 2, and Midjourney. Human evaluators rated the outputs for clarity, imagination, and alignment with the input description.

Table 4: Average Human Evaluation Scores (Scale: 1–10)

Prompt	Stable Diffusion	DALL·E 2	Midjourney
A red apple on a wooden table	9.2	8.4	8.8
A panda skateboarding in Times Square	8.4	8.8	9.3
A surreal city floating in the sky	8.1	8.9	9.4

Stable Diffusion outperformed DALL·E 2 in simpler prompts due to its higher photorealism and balanced contrast. However, Midjourney excelled in abstract prompts with visually imaginative outputs.

5.6 Result analysis

All image generations were used using an Nvidia RTX 3060 GPU (12 GB VRAM). The time of a 512 × 512 image per average generation was about 7.2 seconds. GPU usage increased from 78% to 92% depending on rapid complexity.

The combination of structural (SSIM) and pixel-based (MSE, RMSE) Assessment with a human perception-based score provides a comprehensive approach to model performance. The balance between these matrices suggests that the model produces high quality, structurally consistent images with smaller boundaries in abstract interpretation.

6. ACKNOWLEDGEMENT

We sincerely want to thank our College Guide, Dr. Nabanita Mandal, for her valuable guidance, assistance, and encouragement throughout this process. Her realistic reactions and innovative criticism played a crucial role in shaping our understanding and examining the key points. We also appreciate Thadomal Shahani Engineering College for providing essential resources and a supportive learning environment that helped us complete this assignment. Special method for the Department of Artificial Intelligence and Computer Science to promote the culture of investigation and innovation. We sincerely acknowledge the support and infrastructure provided by Hugging Face and other open-source contributors, whose tools and platforms greatly enhanced our ability to explore the potential of text-to-image diffusion models.

7. CONCLUSION

The Stable Diffusion version marks a significant advancement in the field, offering a promising opportunity for traditional liberal arts fashion as Anthem. Its ability to generate brilliant, diverse snapshots from textual prompts has broad implications for numerous applications, from art and design to marketing and beyond. The open-source nature of the version, facilitated through Hugging Face, has further driven innovation by allowing researchers and developers to customize and expand its capabilities. Our overview highlights the strengths of Stable Diffusion, especially in terms of stability, high-quality images, and computational efficiency, while also recognizing the challenges that remain in AI-driven image generation. As diffusion models continue to evolve, they hold the potential to revolutionize creative industries and push the boundaries of what is possible with AI-generated visual content. Future research should focus on addressing current limitations, including improving the diversity of generated images and reducing computational costs, to further enhance the capabilities of text-to-image synthesis models.

8. REFERENCES

[1] M. Elasri, O. Elharrouss, S. Al-Maadeed, and H. Tairi, “Image generation: A review,” Neural Processing Letters, Feb. 2022. [Online]. Available: https://doi.org/10.1007/s11063-022-10777-x

[2] S. Ramzan, M. M. Iqbal, and T. Kalsum, “Text-to-image generation using deep learning,” Engineering Proceedings, vol. 20, p. 16, Jul. 2022. [Online]. Available: https://doi.org/10.3390/engproc2022020016

[3] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” arXiv preprint arXiv:2102.12092, Feb. 2021. [Online]. Available: https://arxiv.org/abs/2102.12092

[4] Youngsun Lim, Hojun Choi, Hyunjung Shim, “I-HallA: Evaluating Image Hallucination in Text-to-Image Generation with Question Answering,” arXiv preprint arXiv:2409.12784, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2409.12784

[5] Zhongjie Duan, Qianyi Zhao, Cen Chen, Daoyuan Chen, Wenmeng Zhou, Yaliang Li, Yingda Chen, “ArtAug: Enhancing Text-to-Image Generation through Synthesis-Understanding Interaction,” arXiv preprint arXiv:2412.12888v2, Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.12888

[6] Ling Yang, Zhilong Zhang, Zhaochen Yu, Jingwei Liu, Minkai Xu, Stefano Ermon, Bin Cui, “ContextDiff: Contextualized Diffusion Model for Text-to-Image and Text-to-Video Generation,” arXiv preprint arXiv:2402.16627v3, Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.16627

[7] Jiatao Gu, Yuyang Wang, Yizhe Zhang, Qihang Zhang, Dinghuai Zhang, Navdeep Jaitly, Josh Susskind, Shuangfei Zhai, “DART: Denoising Autoregressive Transformer for Scalable Text-to-Image Generation,” arXiv preprint arXiv:2410.08159v2, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.08159

[8] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and

Jie Tang, “ComposeAnything: Compositional Text-to-Image Generation with Layout Reasoning,” arXiv preprint arXiv:2403.05121v1, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2403.05121

[9] Zeeshan Khan Shizhe Chen Cordelia Schmid, “Commonsense-T2I: A Benchmark for Evaluating Commonsense Reasoning in Text-to-Image Models,” arXiv preprint arXiv:2505.24086v1, May 2025. [Online]. Available: https://arxiv.org/abs/2505.24086

[10] Junsong Chen1,2,4, Yue Wu1, Simian Luo3, Enze Xie1†, Sayak Paul5, Ping Luo4, Hang Zhao3, Zhenguo Li1, “PIXART-δ: Fast and Controllable Image Generation with Latent Consistency Models,” arXiv preprint arXiv:2401.05252v1, Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.05252

[11] Ling Yang1∗† Zhilong Zhang1∗ Zhaochen Yu1∗ Jingwei Liu1 Minkai Xu2 Stefano Ermon2 Bin Cui1†, “ContextDiff: Contextualized Diffusion Model for Text-to-Image and Text-to-Video Generation,” arXiv preprint arXiv:2402.16627v3, Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.16627