

Achieving High-Quality Text and Audio-to-Image Generation in a Single Step

Satish Karanjekar¹, Yashwant Tanwar², Shruti Bhagat³, Prof. D.A. Sananse⁴

U.G Students, Department of Computer Science and Engineering^{1,2,3}

Professor, Department of Computer Science and Engineering⁴

Jawaharlal Darda Institute of Engineering and Technology, Yavatmal, Maharashtra, India

Abstract - Diffusion models have significantly advanced text-to-image generation by producing high-quality and imaginative results. However, their multi-step sampling process often proves slow, requiring extensive inference steps to achieve satisfactory outcomes. Despite attempts to improve sampling speed and computational efficiency through distillation, creating a functional one-step model has remained elusive. In this study, we investigate Rectified Flow, a recent method primarily applied to small datasets, as a potential solution. Central to Rectified Flow is its reflow procedure, which optimizes probability flow trajectories, refines noise-to-image mapping, and enables effective distillation with student models. We introduce a novel text-conditioned pipeline to convert Stable Diffusion (SD) into an ultra-fast one-step model. Our approach underscores the crucial role of reflow in enhancing noise-to-image assignments. Leveraging this pipeline, we develop the first one-step diffusion-based text-to-image generator capable of producing high-quality images comparable to those generated by SD. Additionally, we extend our methodology to include audio inputs, demonstrating its efficacy in generating images from audio cues with remarkable fidelity and speed.

Key Words: Stable Diffusion, Text-to-Image Creation, Image Processing

I. INTRODUCTION

Cutting-edge text-to-image (T2I) generative models have taken substantial strides in crafting lifelike and intricate images grounded in textual depictions. Innovations like DALL-E, Imagen, Stable Diffusion, StyleGAN-T, and GigaGAN showcase remarkable prowess in producing varied visual content with fidelity. These strides owe much to extensive datasets and intricate generative designs.[2][3][3]

Nevertheless, despite their impressive output, T2I models grapple with challenges in inference speed and computational resources. This is glaringly apparent in models adopting auto-regressive or diffusion methods, such as Stable Diffusion, necessitating numerous steps for image generation.[3] The computational overhead limits their practicality, particularly where swift image creation is vital.

To surmount these hurdles, various strategies, including knowledge distillation, have emerged.[4] While promising in

reducing inference time, they falter in low-step scenarios, leaving the quest for one-step large-scale diffusion models unresolved.

In this study, we present a pioneering one-step model derived from Stable Diffusion (SD). Propelled by the necessity for swift image synthesis, we expand our methodology to encompass audio inputs, enabling image generation from textual and auditory cues concurrently.[5] However, applying knowledge distillation directly to SD yields subpar results due to noise-to-image alignment issues.

To conquer this obstacle, we harness Rectified Flow, a recent generative breakthrough employing probabilistic flows. Specifically, we employ reflow, a novel technique gradually straightening probability flow trajectories to enhance noise-to-image alignment.[6] This refinement streamlines the distillation process, enabling training of a one-step SD model capable of producing high-quality, detailed images.[6][7]

Our one-step model exhibits state-of-the-art performance on benchmark datasets, surpassing prior methods in image quality and inference speed. Moreover, we extend the model to accommodate audio inputs, showcasing its adaptability and efficacy in multimodal image generation.[8][9][10] Altogether, our work constitutes a significant stride in text and audio-to-image synthesis, furnishing a pragmatic solution for swift, high-fidelity image creation.

II. ANALYSIS OF PROBLEM

Current text-to-image (T2I) generative models have displayed impressive progress in generating images from textual descriptions, but they are not devoid of their shortcomings. One notable obstacle lies in the prolonged inference time and computational consumption linked with many of these models, particularly those employing auto-regressive or diffusion-based methodologies.[10] These models frequently demand numerous inference steps to produce satisfactory outcomes, resulting in substantial computational overhead and impeding real-time or swift image generation.[11]

Furthermore, scalability remains a concern, notably in the realm of one-step large-scale T2I models. Although models such as StyleGAN-T and GigaGAN have attained significant success, they depend on generative adversarial training, requiring meticulous tuning of both the generator and

discriminator. This dependency on elaborate training procedures restricts their scalability and might present challenges when applied to diverse datasets or domains.[12][13]

Another notable issue arises with knowledge distillation methodologies, which strive to diminish sampling steps and hasten inference. Despite their potential, these techniques often encounter difficulties in the small-step regime, complicating the development of efficient one-step large-scale diffusion models. Additionally, endeavours to distil models like Stable Diffusion have faced obstacles due to sub-optimal coupling between noises and images, hindering the distillation process and yielding unsatisfactory outcomes.[14]

Overall, tackling these challenges is pivotal for propelling the field of T2I generative models forward. Overcoming hurdles related to inference time, scalability, and knowledge distillation will pave the path for more streamlined, scalable, and top-notch image synthesis from textual descriptions, unlocking fresh opportunities for applications ranging from content creation to assistive technologies.[15]

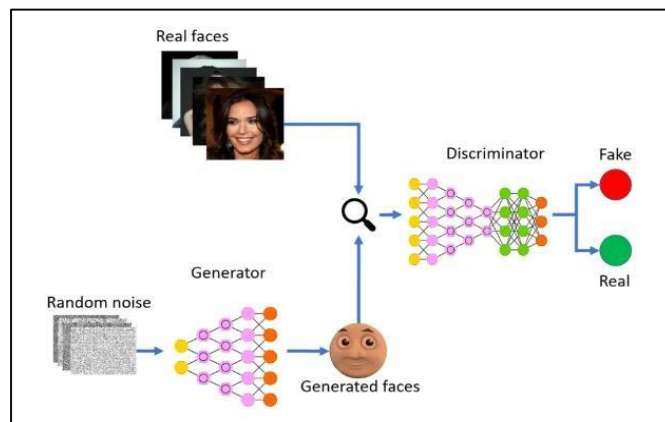


Fig -1: Generative Adversarial Network

III. RELATED WORK

Generative frameworks for the creation of images confront distinct challenges due to the elevated complexity of visuals. Generative Adversarial Networks (GANs) offer efficient exploration of high-resolution images with satisfactory perceptual fidelity; however, they pose challenges in fine-tuning and fail to capture the entire data distribution comprehensively. Conversely, likelihood-based methodologies prioritize precise density estimation, facilitating more compliant optimization.[16] While variational autoencoders (VAE) and flow-based models excel in synthesizing high-resolution images, they lag behind GANs in sample quality. Autoregressive models (ARM) face limitations in resolution due to sequential sampling and computationally intensive designs, despite their proficiency in density estimation. Maximum-likelihood training disproportionately allocates resources to model subtle, high-frequency features present in pixel-based image representations, prolonging training durations.[17] To mitigate these issues, many approaches employ two-stage techniques, compressing latent image spaces

with ARMs to scale to higher resolutions. Recent advancements in sample quality and density estimation have been achieved with Diffusion Probabilistic Models (DM), especially when implemented with UNets, aligning with image-like data biases. However, evaluating and enhancing models in pixel space incur low inference speeds and significant training costs. Our proposed Low-Dimensional Models (LDMs) address these challenges by operating in a compressed latent space of reduced dimensionality. Employing a two-stage strategy, combining different techniques, and leveraging convolutional backbones, we achieve scalability to larger latent spaces without compromising computational efficiency, thereby striking a balance between learning essential features and maintaining high-fidelity reconstructions. Strategies integrating encoding/decoding model pairs offer promising avenues for future exploration, either jointly or independently improving upon previous results.

IV. PROPOSED METHODOLOGY

We observe that notwithstanding the fact that distribution models enable the disregard of perceptually insignificant nuances by inadequately sampling the loss terms that correspond, They still necessitate costly function evaluations in pixel space, which imposes a considerable burden on energy and computation resources. The aim is to diminish the computational load needed to train diffusion models for the creation of high-resolution images. We propose rectifying this issue by explicitly segregating the compressive and generative learning phases.[18] Employing an auto-encoding approach accomplishes this, acquiring a space that closely resembles the image space in terms of perception but with significantly reduced computational complexity.[19] Such an approach yields several advantages: i) By departing from the high-dimensional picture space, we achieve diffusion models that utilize sampling in a low-dimensional space, markedly enhancing computational efficiency. (ii) We leverage the inductive bias inherent in diffusion models; their UNet architecture, inherited [14], renders them particularly suited for spatially structured data without necessitating high compression levels that compromise quality, unlike other techniques. (iii) Finally, we generate versatile compression models that can be utilized in single-image CLIP-guided synthesis [15] and various downstream applications, wherein their latent space can train numerous generative models.

V. SYSTEM DESIGN AND IMPLEMENTATION

V.I Overview

Diffusion Explainer is a tool that helps you understand how Stable Diffusion creates a detailed image from a text prompt that you choose. It shows you step by step how random noise turns into a clear picture, and you can control the speed of this process using the Timestep Controller. There are two main ways to look at this tool: the Architecture View, which gives you an overall look at how Stable Diffusion works and lets you dig deeper if you want, and the Refinement Comparison View, which lets you compare how different text prompts affect the image creation process. This tool is made using basic web

technologies like HTML, CSS, and JavaScript, along with the D3.js library for visualization.[19][20] There are 13 different text prompts you can choose from, all based on a template from a book called "A Traveler's Guide to the Latent Space." These prompts include popular keywords taken from various sources like literature and articles, making it easier for you to explore different aspects of image generation.

V.II Architecture View

The Architecture View gives an overview of how the Text Representation Generator changes a written prompt into vector representations that direct the Image Representation Refiner to gradually refine random noise into a detailed image representation matching the prompt. By clicking on the generators, users can delve into their inner workings.

1. Text Representation Generator

This generator transforms text prompts into vector representations. Clicking on it reveals the Text Operation View, which details how the prompt is broken down into tokens by the Tokenizer and how the Text Encoder converts these tokens into vector representations.

Clicking on the Text Encoder reveals the Text-image Linkage Explanation, which visually shows how Stable Diffusion connects text and image by utilizing the CLIP text encoder to create text representations with image-related information.

2. Image Representation Refiner

The Image Representation Refiner gradually refines random noise into a detailed image representation that matches the input text prompt. The Diffusion Explainer shows the image representation at each refinement step in two ways: (1) as a small image using linear operations and (2) upscaled to Stable Diffusion's output resolution. Users can access the Image Operation View by expanding the Image Representation Refiner, which explains how the UNet neural network predicts and weakens noise in the image representation to better match the prompt. The guidance scale hyperparameter, controlling how closely the image adheres to the text prompt, is explained at the bottom and further clarified in the Interactive Guidance Explanation through a slider, enabling users to experiment with different values and understand how higher values increase adherence between the generated image and the text prompt.

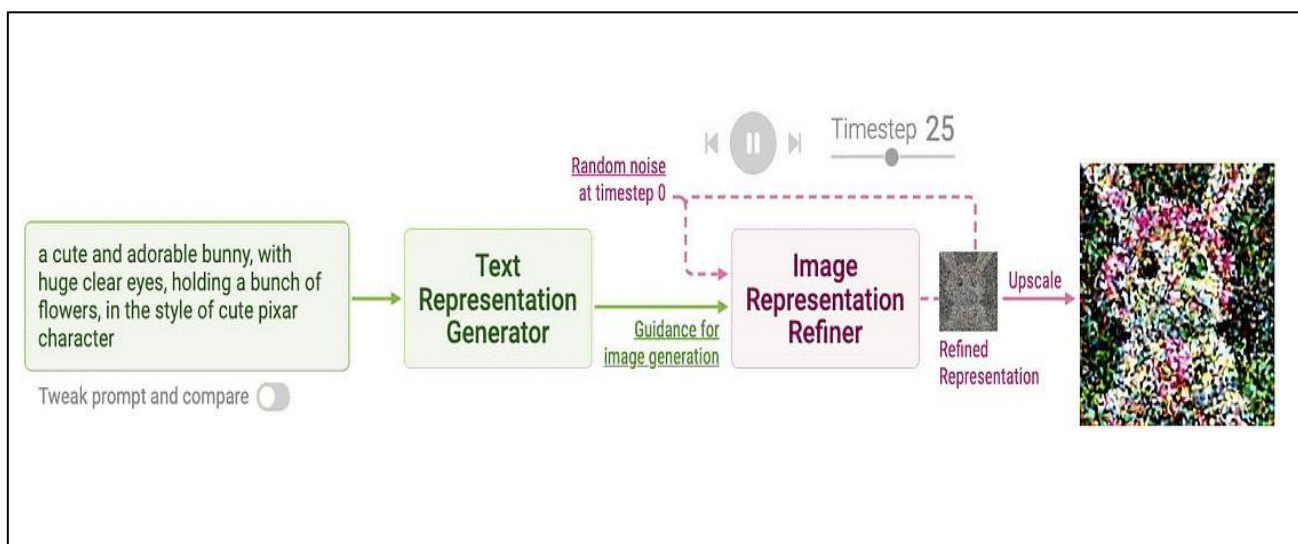


Fig -2: Stable Diffusion Explained Step-by-Step with Visualization

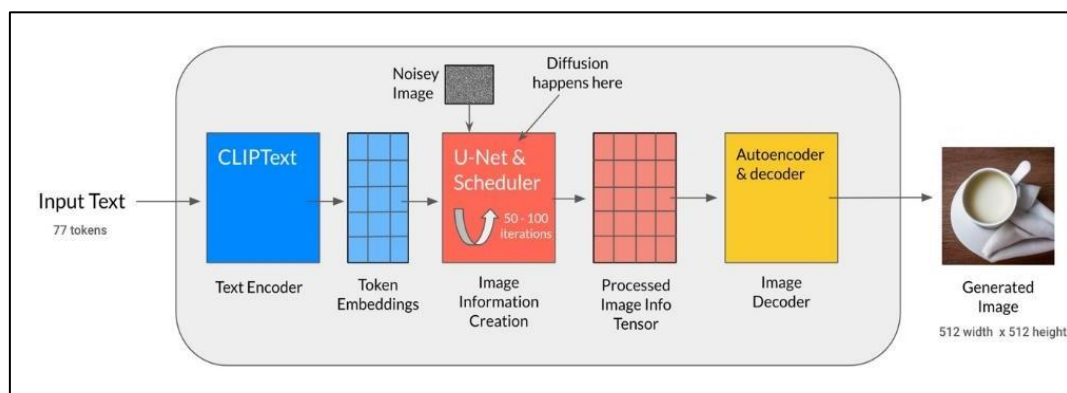


Fig -3: Stable Diffusion Architecture

VI. EXPERIMENTAL RESULT

This section shows the result of our model in terms of inception score. It is found that the proposed model performs well.

The below table shows the inception score of previously used model and our proposed model. Based on the score diffusion model works effectively.

Ref.	Model	Inception Score (IS)
Model-1	StackGAN	3.21± 0.03
Model-2	StackGAN++	3.25±0.02
Model-3	HDGAN	3.47± 0.06
Our Proposed Method	LDM using stable diffusion	5.2± 0.05

Table -1: Models and Inception Score

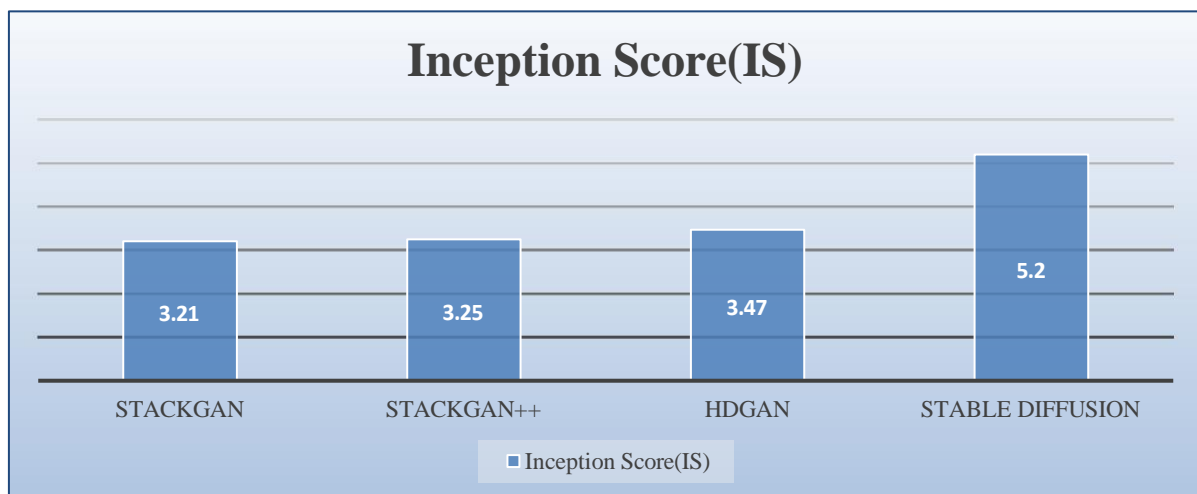



Chart-1. Bar Chart of Inception Score

TEXT/AUDIO (input)	Output
Dog with Cat	

Rainbow above clouds	   
Girl dancing in rain	   
Bike riders on roads	   

Table -2: Sample Output

VII. CONCLUSION AND FUTURE SCOPE

Creating good images from written descriptions is a fascinating area of study with many practical uses. However, it's quite hard because real-world language and visual descriptions are messy and vary a lot. Most methods for turning text into images right now aim to make images as a whole, not paying much attention to what's in the front and what's in the back. This can make objects in the images blend into the background too much. Also, these methods often don't consider how different types of models that create things can work together. We can make denoising diffusion models better at training and making samples without making them worse by using latent diffusion models, which are a quick and simple way to do it. If we don't use specific designs based on this and our way of paying attention to different parts of the text, our research might do better than what's being used now for making pictures based on conditions. Even though making pictures with latent diffusion models is

slower than with other methods, like GANs, and even though it needs less computer power than methods that focus on pixels, our models don't lose much picture quality. However, when we try to make the pictures really detailed, our models might not be as good. We think this is a place where our models for making pictures clearer are limited. But we can make things better, as we've shown in our results. We need to make sure our pictures match the quality of the text, and we can also make people better at understanding pictures.

VIII. REFERENCES

- 1] J. Alammar. The illustrated Stable Diffusion. <https://jalammar.github.io/illustrated-stable-diffusion/>, 2022. Accessed on: 2023-04-30.
- 2] J. Alammar. AI Art Explained: How AI Generates Images (Stable Diffusion, Midjourney, and DALL-E). <https://youtu.be/MXmacOUJUaw>, 2023. Accessed on: 2023-04-30.
- 3] Andrew. Absolute beginners guide to Stable Diffusion AI image. <https://stable-diffusion-art.com/beginners-guide/>, 2023. Accessed on: 2023-04-30.
- 4] Andrew. How does Stable Diffusion work? <https://stable-diffusion-art.com/how-stable-diffusion-work/>, 2023. Accessed on: 2023-04-30.
- 5] Andrew. Stable Diffusion prompt: a definitive guide. <https://stablediffusion-art.com/prompt-guide/>, 2023. Accessed on: 2023-04-29.
- 6] anton. Announcing Stable Attribution - A tool which lets anyone find the human creators behind AI generated images. <https://twitter.com/atroy/n/status/1622355473193381888>, 2023. Accessed on: 2023-04-30.
- 7] J. Brusseau. Acceleration AI Ethics, the Debate between Innovation and Safety, and Stability AI's Diffusion versus OpenAI's Dall-E. arXiv preprint arXiv:2212.01834, 2022.
- 8] Engler. Early thoughts on regulating generative AI like ChatGPT. Brookings Institution, 2023. Accessed on: 2023-04-30.
- 9] Y. Hosni. Getting Started With Stable Diffusion. <https://medium.com/towards-artificial-intelligence/getting-started-withstable-diffusion-f343639e4931>, 2022. Accessed on: 2023-04-30.
- 10] J. Howard. From Deep Learning Foundations to Stable Diffusion. <https://www.fast.ai/posts/part2-2023.html>, 2023. Accessed on: 2023-04-30.
- 11] J. Huber and A. Troynikov. Stable Attribution. <https://www.stableattribution.com>, 2023. Accessed on: 2023-04-30.
- 12] V. Liu and L. B. Chilton. Design Guidelines for Prompt Engineering Text-to-image Generative Models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pages 1–23, 2022.
- 13] Q. Nichol and P. Dhariwal. Improved Denoising Diffusion Probabilistic Models. In International Conference on Machine Learning, pages 8162–8171. PMLR, 2021.
- 14] OpenAI. DALL-E 2. <https://openai.com/product/dall-e-2>, 2022. Accessed on: 2022-09-28.
- 15] J. Oppenlaender. A Taxonomy of Prompt Modifiers for Text-to-Image Generation. arXiv preprint arXiv:2204.13988, 2022.
- 16] S. Patil, P. Cuenca, N. Lambert, and P. v. Platen. Stable Diffusion with Diffusers. <https://huggingface.co/blog/stable-diffusion>, 2022. Accessed on: 2023-04-30.
- 17] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. CoRR, abs/2107.00630, 2021.
- 18] Kevin Frans, Lisa B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language image encoders. ArXiv, abs/2106.14843, 2021.
- 19] Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. arXiv 2016, arXiv:1605.05396.
- 20] Zhang, Z.; Xie, Y.; Yang, L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6199–6208.
- 21] Cai, Y.; Wang, X.; Yu, Z.; Li, F.; Xu, P.; Li, Y.; Li, L. Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network. IEEE Access 2019, 7, 183706–183716