INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 3, MARCH 2026 | ISSN: 2456-4184 | IJNRD.ORG

Efficient Local Text-to-Image Generation Using Stable Diffusion XL with the Fooocus Interface

[1]AARY JADHAV, [2]YASHAS MAYEKAR, [3]PUSHPAK SHINDE, [4]NASIR ZAIDI, [5*]ASHWINI SANKHE
[1] Artificial Intelligence & Machine Learning, Universal College of Engineering/Mumbai University, Vasai, India
[2] Artificial Intelligence & Machine Learning, Universal College of Engineering/Mumbai University, Vasai, India
[3] Artificial Intelligence & Machine Learning, Universal College of Engineering/Mumbai University, Vasai, India
[4] Artificial Intelligence & Machine Learning, Universal College of Engineering/Mumbai University, Vasai, India
[5*] Artificial Intelligence & Machine Learning, Universal College of Engineering/Mumbai University, Vasai, India

Abstract: With the development of advanced generative artificial intelligence technology, it is now possible to create high-quality images from text prompts using diffusion-based generative models. Stable Diffusion is a strong open-source tool that allows for the efficient generation of images using latent diffusion models. However, the deployment of these models is often complicated. In this paper, a simple text-to-image generation system is proposed using the Stable Diffusion XL model with the Fooocus interface. In the proposed method, efficient image generation is possible locally, with minimal complexity in the configuration of the system. In the proposed system, the prompt is first encoded using a CLIP text encoder, then the latent diffusion process is carried out using a U-Net model, and finally, the image is reconstructed using a Variational Autoencoder. From the experimental results, it is clear that the proposed system is efficient in the generation of high-quality images from text prompts.

Keywords **- Generative AI, Stable Diffusion, Diffusion Models, Text-to-Image Generation, Latent Diffusion Models**

I. INTRODUCTION

Generative Artificial Intelligence is one of the most significant technologies to have emerged in recent years. Generative Artificial Intelligence is a branch of modern machine learning and computer vision. Recent developments in deep generative models have enabled computers to generate images, audio, and videos from text descriptions. Text-to-image generation is a significant application of this field, as computers can generate images from text. The application of text-to-image generation is vast, ranging from digital art, prototyping, video games, advertisement, and data augmentation. The initial text-to-image model was developed using Generative Adversarial Networks. The Generative Adversarial Network model uses adversarial training to generate images. The model was successful, but there were some problems with training the model, which led researchers to look for other alternatives. The Generative Adversarial Network model was not effective in generating images, so a new model was developed, known as the diffusion model. The diffusion model was developed to generate images from text. The model generates images through a process known as iterative denoising. The model was successful, and some popular models developed using this approach are DALL-E 2, Imagen, and Stable Diffusion.

The model of Stable Diffusion has received considerable attention because it is open-source and works effectively on consumer-grade hardware. Unlike the previously used diffusion models, which work directly in pixel space, Stable Diffusion employs latent diffusion models that work on the compressed version of the image. This reduces the computational requirements of the model significantly. Despite the advantages of the Stable Diffusion model, it is still a complex task for users to work effectively with the model. Most models expose a number of parameters, which can be overwhelming for the user. To address the issue of usability, user-friendly interfaces like Fooocus have been developed. This research aims to develop a simplified text-to-image generation pipeline using the Stable Diffusion XL model with the help of the Fooocus interface. This will help users understand the usability of the Stable Diffusion model and the generation of high-quality images using the model.

The contributions of the proposed work include:
1. Implementation of a simplified text-to-image generation pipeline using the Stable Diffusion XL model.
2. Integration of the Stable Diffusion XL model with the Fooocus interface.
3. Analysis of the diffusion-based generation pipeline.
4. Evaluation of the system's effectiveness in generating high-quality images from a variety of text prompts.

II. PROPOSED SYSTEM

II.I. GENERATIVE ARTIFICIAL INTELLIGENCE

Generative artificial intelligence refers to machine learning models capable of generating new content that resembles data from a training distribution. These models learn underlying patterns within datasets and use this knowledge to produce new outputs. Generative models are widely used in applications such as image generation, natural language processing, music generation, and simulation environments. Two major categories of generative models include GAN-based models and diffusion-based models. GANs use adversarial training between generator and discriminator networks, while diffusion models rely on probabilistic noise processes to generate data.

II. II. DIFFUSION MODELS

Diffusion models generate images by learning to reverse a gradual noise-adding process. During training, noise is progressively added to images until they become random noise. The model then learns to reverse this process by predicting and removing noise

step by step. The generation process begins with random noise and iteratively removes noise through multiple denoising steps until a coherent image emerges. This approach allows diffusion models to produce high-quality and diverse outputs.
Mathematically, diffusion models involve two processes:

i. Forward Diffusion Process
Gradually adds Gaussian noise to an image.

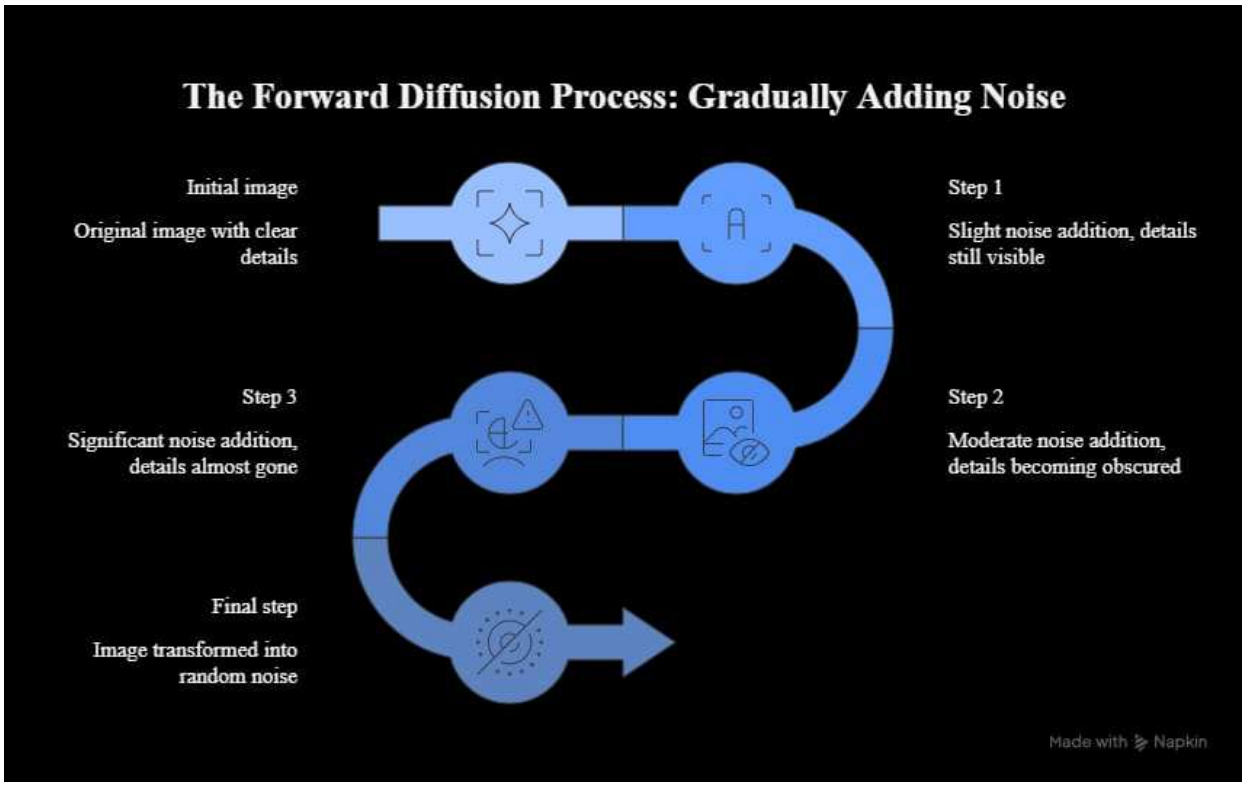

Figure 2.1: Forward process of the diffusion model where Gaussian noise is progressively added to the data over multiple timesteps.

ii. Reverse Diffusion Process
A neural network predicts noise and removes it to reconstruct the image.

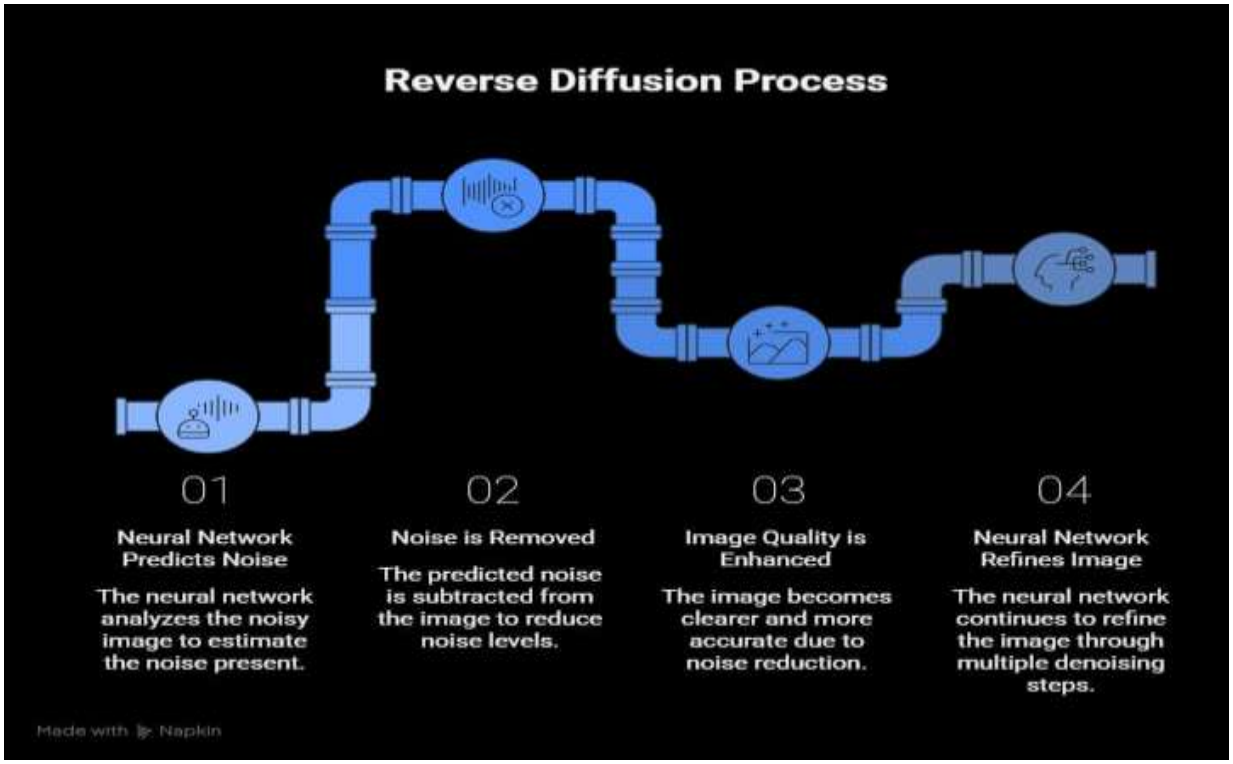

Figure 2.2: Reverse process of the diffusion model where noise is progressively removed to reconstruct the image.

II. III. Latent Diffusion Models

Directly operating the diffusion process in the pixel space requires large computational resources. This challenge is solved by the Latent Diffusion Models (LDMs). The LDMs solve the challenge by performing the diffusion process in the compressed latent space. The process works as follows: The images are encoded into the latent space by the Variational Autoencoder (VAE). The process then continues in the latent space. After the denoising process, the images are decoded to the final images.

III. Literature Review

Name: Empowering Local Image Generation: Harnessing Stable Diffusion for Machine Learning and AI
Authors: Ahmed Imran Kabir, Limon Mahomud, Abdullah Al Fahad, Ridwan Ahmed
Proposed Work: A practical guide for deploying Stable Diffusion locally via AUTOMATIC1111 WebUI on consumer hardware, covering SDXL 1.0, LoRA, ControlNet, prompt engineering, and a SWOT analysis of local AI image generation.
Difference: Kabir et al. use AUTOMATIC1111 which exposes complex controls for technical users, whereas Fooocus prioritises simplicity through smart defaults and a streamlined interface, making it accessible to non-technical users with the same SDXL foundation.

Name: Denoising Diffusion Probabilistic Models (DDPM)
Authors: Jonathan Ho, Ajay Jain, Pieter Abbeel
Proposed Work: Established the theoretical foundation for diffusion models by introducing the DDPM framework. Demonstrated that diffusion models can match GAN-level quality through a noise prediction training objective, achieving a state-of-the-art FID of 3.17 on CIFAR-10.
Difference: DDPM operates in pixel space on small resolutions and performs unconditional generation with no text input. Fooocus builds on this theory but operates in latent space via SDXL, supports text-to-image generation, and is optimised for high-resolution 1024×1024 output on consumer hardware.

Name: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis
Authors: Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, Robin Rombach
Proposed Work: Introduced SDXL with a 2.6B parameter U-Net, dual CLIP text encoders, and three micro-conditioning innovations — size conditioning, crop conditioning, and multi-aspect ratio training — along with an optional refinement model achieving 48.44% user preference over prior versions.
Difference: SDXL is the core model Fooocus is built upon. Fooocus extends it by adding optimised samplers, automatic prompt enhancement, LoRA integration, a comprehensive style system, and a simplified user interface — none of which are part of the original SDXL paper.

Name: RAPHAEL: Text-to-Image Generation via Large Mixture of Diffusion Paths
Authors: Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, Ping Luo
Proposed Work: Introduced a Mixture-of-Experts based diffusion model with space-MoE and time-MoE layers creating billions of diffusion paths. Also proposed edge-supervised learning for improved detail. Achieved a state-of-the-art FID of 6.61 on MS-COCO and outperformed all competitors in human evaluation on ViLG-300.
Difference: RAPHAEL is a research-focused model requiring 1,000 A100 GPUs for training and is not openly available for local deployment. Fooocus uses SDXL which is open-source and consumer hardware compatible, prioritising practical accessibility over raw benchmark performance.

Table 3.1: Literature survey

Publication Year	Paper Title	Author	Proposed Work	Research Gap
2020	Denoising Diffusion Probabilistic Models (DDPM)	Jonathan Ho, Ajay Jain, Pieter Abbeel	Introduced the DDPM framework for diffusion-based image generation.	Limited to pixel-space, low-resolution, unconditional generation.
2023	SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis	Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, Robin Rombach	Presented SDXL for high-resolution latent diffusion image synthesis.	Focuses on model design, not usability or simplified deployment.

2024	Empowering Local Image Generation: Harnessing Stable Diffusion for Machine Learning and AI	Ahmed Imran Kabir, Limon Mohomud, Abdullah Al Fahad, Ridwan Ahmed	Practical guide for running Stable Diffusion locally with AUTOMATIC1111, covering SDXL, LoRA, ControlNet, and prompting.	Uses a complex interface; lacks focus on beginner-friendly simplicity.
2024	RAPHAEL: Text-to-Image Generation via Large Mixture of Diffusion Paths	Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, Ping Luo	Proposed a large MoE-based diffusion model for high-quality text-to-image generation.	Requires massive resources; not practical for local consumer use.

IV. System Methodology

IV.I. System Overview

The proposed system will follow a text-to-image generation system with a pipeline of Stable Diffusion XL and Fooocus interface. The system will transform text information into images with a multi-stage pipeline of text encoding, latent diffusion, and image reconstruction.

The system architecture is illustrated below:

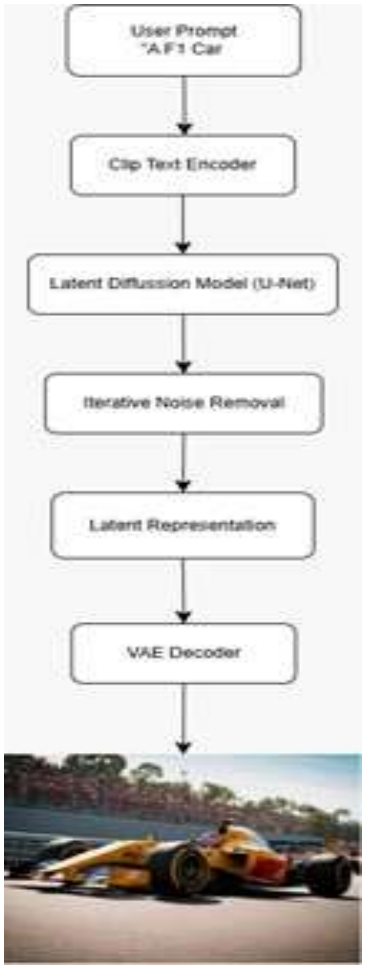

Figure 4.1: Architecture of the Proposed Text-to-Image Generation System

Initially, the system will follow a prompt from a user-defined text information, like “A F1 Car,” which will be taken as an input to generate an image. The text information will be encoded with a text encoder like CLIP to transform it into a high-dimensional embedding vector. This embedding will carry information about the text and will be used to transform it into an image. The text information will be passed to a latent diffusion model with a U-Net structure. In comparison to other traditional diffusion models, which work in the pixel space, the proposed system uses a compressed latent space for the diffusion process. The complexity is reduced, and the important features are preserved. The latent diffusion model uses two input vectors. The first is a randomly initialized noise vector, and the second is the text embedding obtained from the text encoder.

The image generation process is done through an iterative process known as the denoising process. The model starts from random noise, and then the noise is progressively removed. At every time step, the U-Net predicts the noise component that is

present in the latent representation, and then the latent image is updated accordingly. The process continues, and the noisy latent representation is mapped to a structured and semantically meaningful latent image.

Once the process of denoising is over, the resulting feature would have the compressed version of the generated image. However, the resulting feature is not directly interpretable. Hence, the resulting feature is passed through a Variational Autoencoder (VAE) decoder. The VAE decoder decompresses the image from the resulting feature and reconstructs the image in the pixel space.
The final output of the proposed system would be a visually coherent image that represents the input textual prompt. By using the diffusion process in the latent space and the pretrained models like the CLIP encoder and the VAE decoder, the proposed system would be efficient in the image generation process.

IV.II. Prompt Processing

The generation process begins with a user-provided textual prompt. The prompt describes the desired image content, style, or scene. Examples of prompts include:
"A futuristic city skyline at sunset"
"A realistic portrait of a warrior in medieval armor"
"A cyberpunk robot standing in a neon city"

The prompt is processed by the CLIP text encoder, which converts textual input into numerical embeddings. These embeddings represent semantic information about the prompt and guide the image generation process.

IV.III. Diffusion Model

The diffusion model in Fooocus is built upon Stable Diffusion XL (SDXL), which operates through two core processes. In the forward process, Gaussian noise is progressively added to an image over a series of timesteps until the original image is completely destroyed into pure noise. In the reverse process, a 2.6 billion parameter U-Net denoising network learns to iteratively remove this noise step by step, guided by the text prompt provided by the user, ultimately reconstructing a coherent high-quality image from random noise. The U-Net architecture employs cross-attention mechanisms to integrate text conditioning into the denoising process, where query vectors are derived from image features and key-value vectors come from the text embeddings produced by the dual CLIP encoder system. Fooocus implements the DPM++ 2M Karras sampler as its default sampling algorithm, which uses a second-order multistep formulation with a Karras noise schedule to achieve high-quality outputs in 30 to 60 steps, significantly reducing the number of function evaluations required compared to standard DDPM sampling. Classifier-free guidance is applied during inference to improve prompt adherence, combining conditional and unconditional noise predictions with a guidance scale typically set between 4 and 12 within Fooocus. The system also supports multiple speed modes, where the Speed mode uses 30 sampling steps, Quality mode uses 60 steps, and Extreme Speed mode reduces to just 8 steps using LCM-LoRA for near real-time generation.

IV.IV. Variational Autoencoder

The Variational Autoencoder in Fooocus serves as the compression and reconstruction bridge between pixel space and the latent space in which the diffusion process operates. Rather than performing denoising directly on high-resolution pixel data, Fooocus leverages the VAE encoder to first compress an input image from its original pixel dimensions into a compact latent representation, specifically reducing a 1024×1024×3 pixel image down to a 128×128×4 latent tensor, achieving a significant reduction in computational cost while preserving the perceptually meaningful semantic content of the image. The diffusion model then performs all denoising operations within this latent space, making the generation process far more efficient than pixel-space diffusion approaches. Once the denoising process is complete, the VAE decoder translates the final denoised latent representation back into a full-resolution RGB image in pixel space. The VAE used in Fooocus is the improved SDXL autoencoder, which was retrained from scratch with a larger batch size of 256 and exponential moving average weight tracking, delivering superior reconstruction quality with a PSNR of 24.7 and SSIM of 0.73 compared to earlier Stable Diffusion VAE versions. This improved reconstruction quality directly contributes to sharper details, more accurate colour reproduction, and reduced artefacts in the final generated images produced by Fooocus.

IV.V. Fooocus Interface

Fooocus serves as the user interface for interacting with the diffusion pipeline. The interface simplifies the generation process by automatically configuring parameters such as sampling steps, guidance scale, and image resolution.
Users only need to provide a text prompt, making the system accessible to individuals without deep technical knowledge.

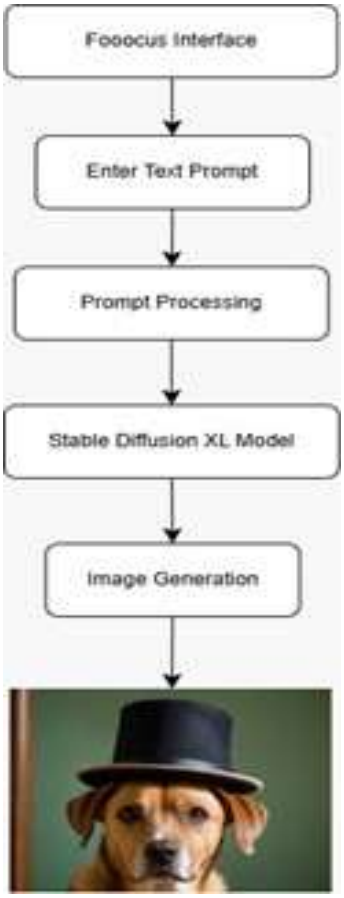

Figure 4.2: Workflow of Image Generation Using Fooocus Interface

The process begins with the Fooocus interface, which serves as the front-end platform for user interaction. Fooocus provides a user-friendly environment that abstracts the complexity of configuring diffusion models, enabling users to generate images without requiring extensive technical knowledge. The user initiates the process by entering a textual prompt through the interface. This prompt describes the desired image content and serves as the primary input to the system. The flexibility of natural language input allows users to specify objects, styles, and contextual details for image generation. Once the prompt is entered, it undergoes a prompt processing stage. In this stage, the input text is refined and optimized to improve image generation quality. This may include tokenization, prompt enhancement, weighting of keywords, and internal formatting to align with the requirements of the Stable Diffusion XL model. Fooocus may also apply default enhancements such as style adjustments and resolution settings to improve output quality.

The processed prompt is then passed to the Stable Diffusion XL (SDXL) model, which acts as the core generative component of the system. SDXL is a latent diffusion model that generates images by iteratively denoising a random latent representation conditioned on the input prompt. It leverages an advanced U-Net architecture, cross-attention mechanisms, and multiple text encoders to ensure high-quality image synthesis and accurate alignment with the textual description. Following model execution, the image generation stage produces the final output. The model transforms the processed prompt into a high-resolution image through a sequence of denoising steps in latent space, followed by decoding into pixel space. The resulting image reflects the semantic content of the input prompt with enhanced visual realism and detail. The integration of Fooocus with the Stable Diffusion XL model significantly reduces the complexity of deploying and using diffusion-based generative models. By automating prompt optimization and system configuration, the proposed pipeline enables efficient, accessible, and high-quality image generation suitable for both research and practical applications.

V. RESULT AND ANALYSIS

The system successfully generated high-quality images across multiple prompt categories. Generated images demonstrated strong alignment with textual descriptions and high visual fidelity.

1. Dashboard

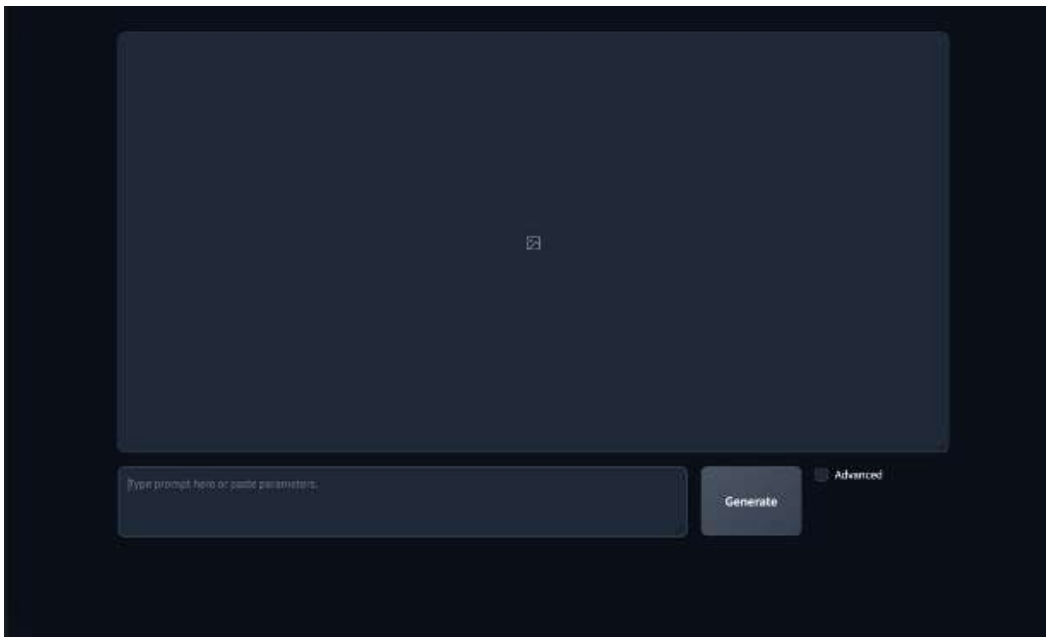

Figure 5.1: Dashboard

The image shows Fooocus's main interface — a dark-themed UI with a large empty canvas where generated images will appear. Below it is a text input box for entering prompts, a Generate button to trigger image generation, and an Advanced toggle for accessing additional settings.

2. Settings Tab

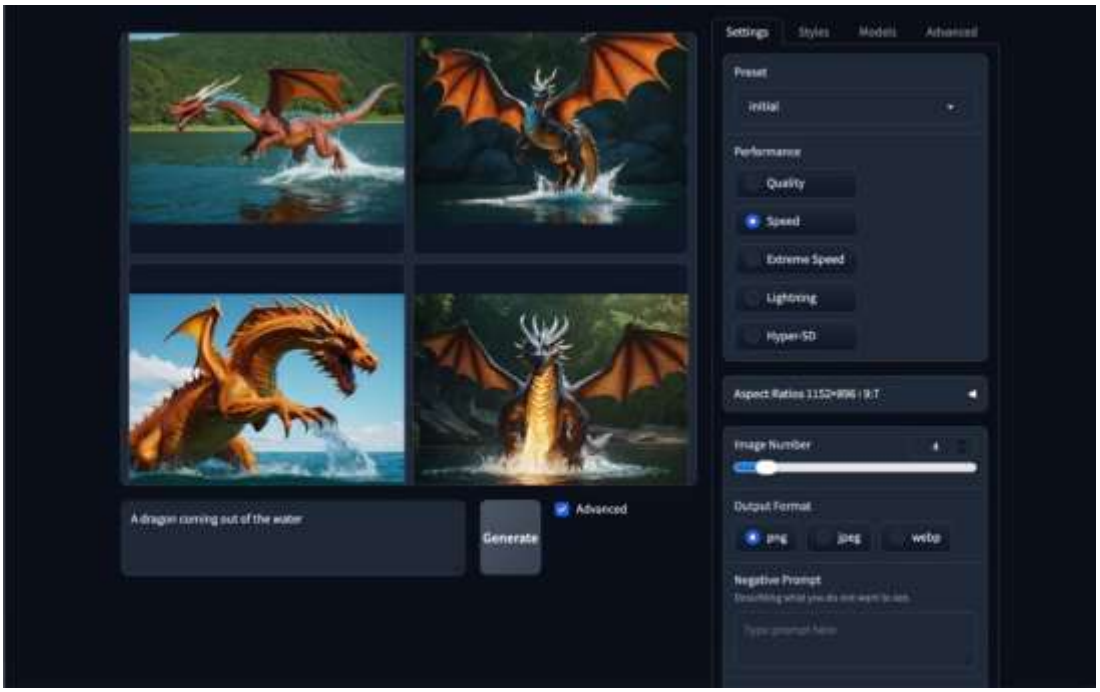

Figure 5.2: Output image generated after configuring with image count 4 and performance mode set to speed.

The Fooocus interface generating four detailed dragon images from a single prompt — "A dragon coming out of the water." The clean and minimal interface allows users to control key parameters such as performance mode, aspect ratio, image count, and output format from the settings panel on the right, highlighting how Fooocus makes high-quality AI image generation straightforward and accessible with minimal user effort.

3. Advance Tab

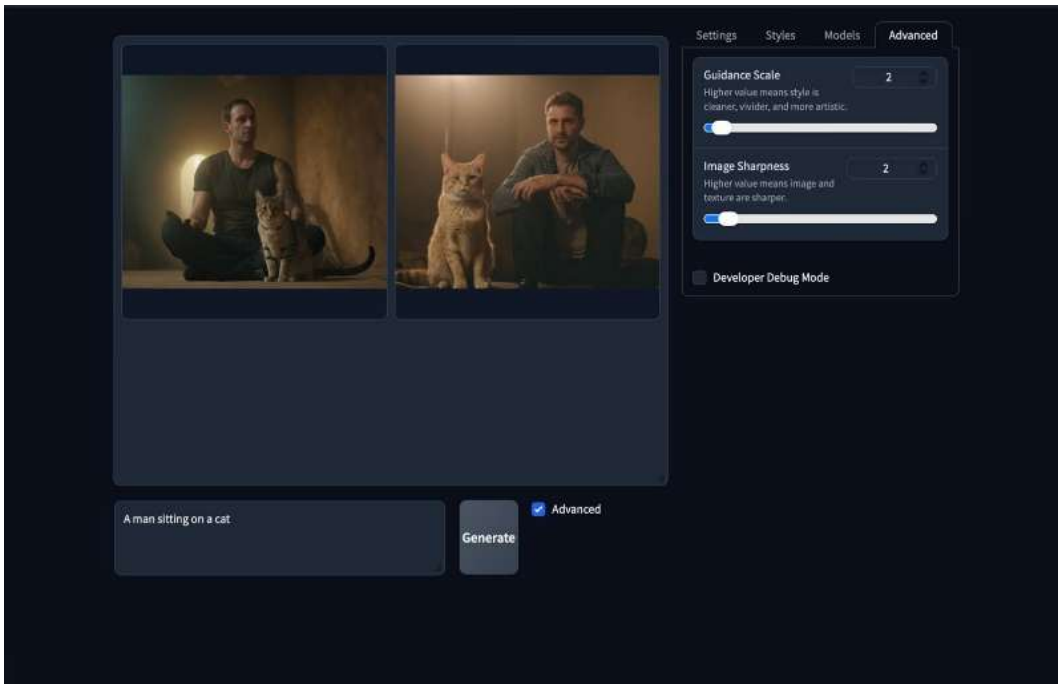

Figure 5.3: Output image generated for guidance scale set to 2 and sharpness set to 2.

Generated from the prompt "A man sitting on a cat" with a low Guidance Scale of 2 and high Sharpness of 13. The elevated sharpness, a feature developed by Fooocus to address SDXL's tendency to produce overly smooth or plastic-looking images, adds visible texture and detail to the output without disturbing the overall composition or structure. The result is a naturally lit, realistic scene with well-defined textures and a stable, clean generation.

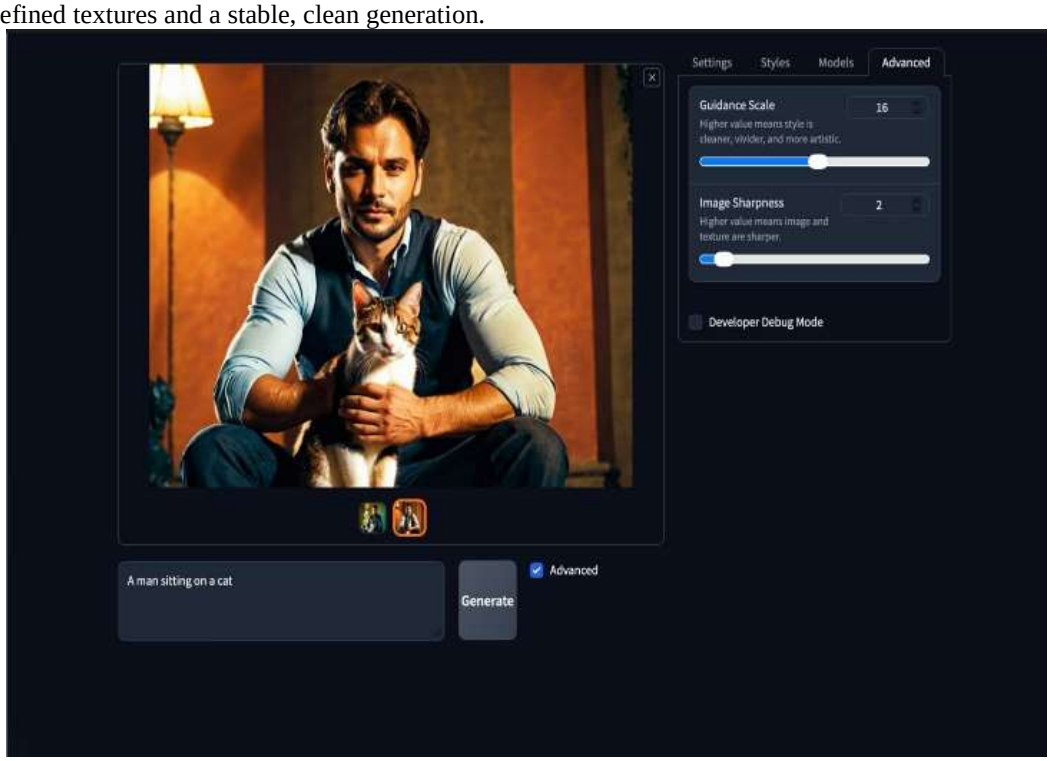

Figure 5.4: Output image generated for guidance scale set to 16 and sharpness set to 2.

The same prompt with Guidance Scale raised to 16 and Sharpness kept low at 2. Fooocus applies adaptive TSNR rescaling on top of the standard CFG scale, preventing the image from appearing burned or oversaturated even at this high guidance value. The output is vivid, cinematic, and strongly prompt-adherent, producing a warmer and more dramatic composition without the typical artefacts that high CFG values usually introduce.

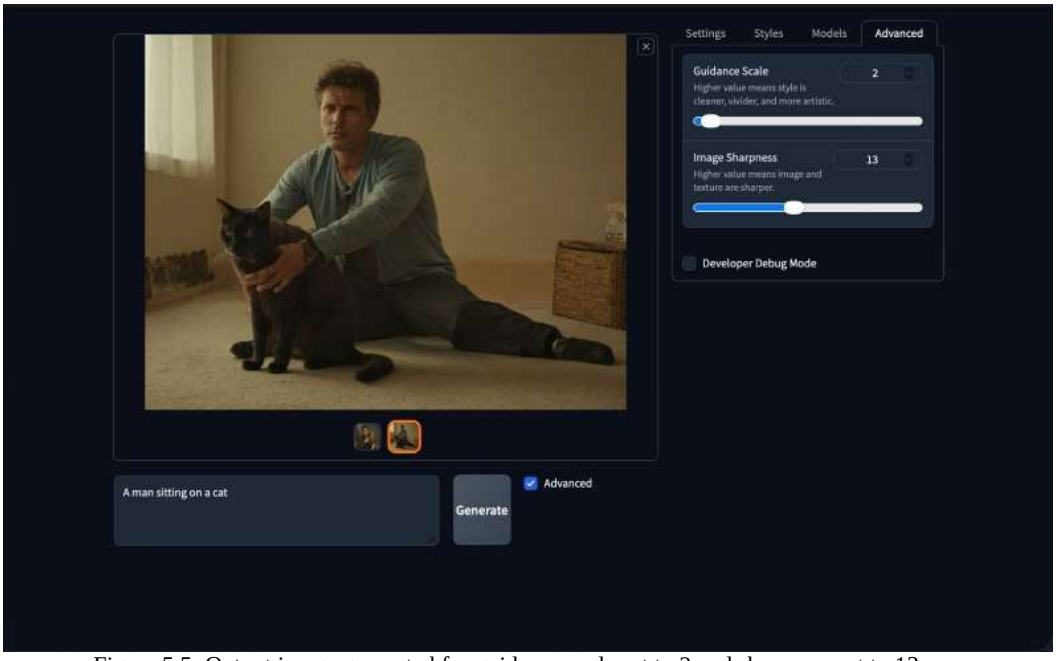

Figure 5.5: Output image generated for guidance scale set to 2 and sharpness set to 13.

Guidance Scale at 2 with Sharpness at 13. At this low guidance value, Fooocus keeps the generation relaxed and interpretive, allowing the model to breathe naturally rather than force prompt adherence. The result is a soft, moody, and cinematically lit composition with organic textures and no over-sharpening artifacts, feeling closer to a candid editorial portrait

4. Styles Tab

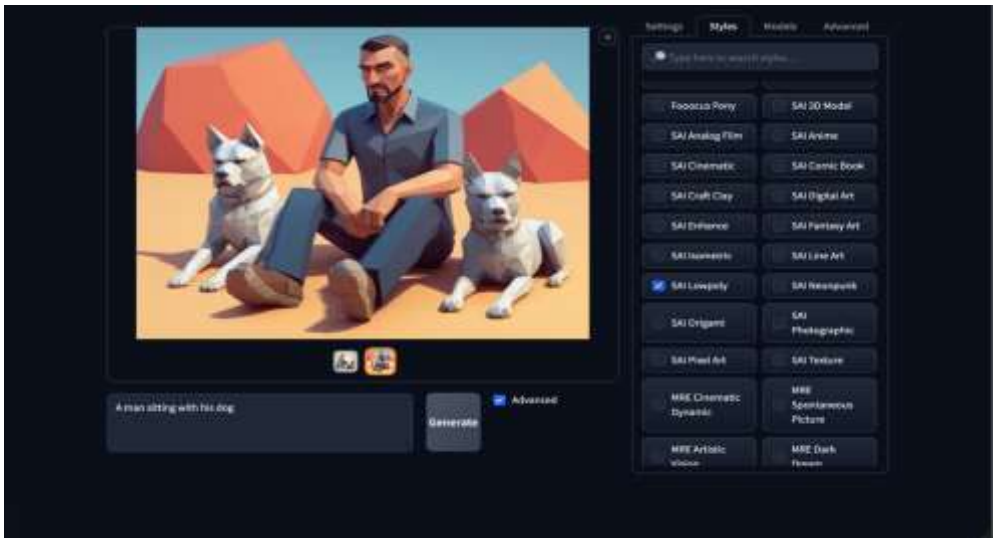

Figure 5.7: Output image generated for selected style SAI Lowpolly.

Fooocus's built-in style system where the prompt "A man sitting with his dog" is generated using the SAI Lowpoly style, producing a distinct geometric low-polygon aesthetic without any additional prompt engineering. The styles panel offers a wide range of options including Cinematic, Anime, Comic Book, Pixel Art, and Fantasy Art, allowing users to completely transform the visual character of their output with a single selection.

VI. Conclusion

This paper presented a simplified implementation of a text-to-image generation system using Stable Diffusion XL integrated with the Fooocus interface. The system demonstrates how advanced diffusion-based generative models can be deployed locally while maintaining usability and high-quality output generation. The study analysed the architecture of diffusion models including text encoding, latent diffusion, and image decoding. Experimental results showed that the system can generate realistic images from diverse textual prompts while maintaining computational efficiency. Future research directions include improving prompt understanding, optimizing generation speed, and expanding the system to support additional generative tasks such as video and 3D content generation.

REFERENCES

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[2] Amarnath Singh, Sanskriti Srivastava, and Ruchi. Text To Image Generation Model: An Efficient Pipeline with Gradio Interface. *IJSREM*, 2023.

[3] Syed sha alam.A, Jeyamurugan.N, Mohamed faiz ali.B, and Veerasundari.R. Stable Diffusion Text-Image Generation. *IJSREM*, 7(02), 2023.

[4] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. RAPHAEL: Text-to-Image Generation via Large Mixture of Diffusion Paths. In *Advances in Neural Information Processing Systems*, 36, 2023.

[5] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation. *arXiv preprint arXiv:2403.12015*, 2024.

[6] Jaineet Shah, Michael Gromis, and Rickston Pinto. Enhancing Diffusion Models for High-Quality Image Generation. Technical Report, 2024.

[7] Zhiyu Tan, Mengping Yang, Luozheng Qin, Hao Yang, Ye Qian, Qiang Zhou, Cheng Zhang, and Hao Li. An Empirical Study and Analysis of Text-to-Image Generation Using Large Language Model-Powered Textual Representation. *arXiv preprint arXiv:2405.12914*, 2024.

[8] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, In So Kweon, and Junmo Kim. Text-to-image Diffusion Models in Generative AI: A Survey. *arXiv preprint arXiv:2303.07909*, 2024.

[9] C. Beulah Christalin Latha, S. V. Evangelin Sonia, G. Linda Rose, Ben M. Jebin, Christhya Joseph, and G. Naveen Sundar. An Efficiency-Optimized Framework for Sequential Image Generation with Stable Diffusion Models. *Journal of Neonatal Surgery*, 14(6s):497–504, 2025.