Stable Diffusion 3 Inpainting, Text-to-Image, and Image-to-Image Generation in Keras Hub

Akif Islam
Department of Computer Science & Engineering
University of Rajshahi
Student ID: 1910776135
Email: iamakifislam@gmail.com

***Abstract*—This report presents a practical study of the Stable Diffusion 3 (SD3) family of models exposed through Keras Hub, focusing on three use cases: end-to-end inpainting, prompt flexibility analysis, and image-to-image and text-to-image generation. An RTX 4090 GPU with the TensorFlow backend was used to run the official SD3 backbone, preprocessor, and task-specific wrappers. Prompt flexibility was evaluated by keeping the input image and mask fixed while systematically changing the prompt semantics and sampling parameters. In addition, the SD3 Text-to-Image and Image-to-Image models were applied to generate diverse scenes from text alone and to stylize a author's own portrait. Qualitative comparison shows that SD3 preserves global structure while allowing strong stylistic changes, with inpainting providing the most localized control, image-to-image offering coherent global transformations, and text-to-image enabling unconstrained synthesis. The study highlights the strengths and limitations of SD3 as a controllable generative model for creative visual tasks and suggests guidelines for prompt design and parameter selection in practical computer vision workflows.**

I. Introduction

Diffusion models have rapidly become the dominant paradigm for high-fidelity image synthesis and editing. Systems such as DALL-E, Midjourney, and the Stable Diffusion family have made it almost routine to turn a short text description into a detailed image. Recent architectures like Stable Diffusion 3 combine latent diffusion with advanced text encoders and cross-attention mechanisms, enabling fine-grained conditioning on natural-language prompts while still remaining computationally efficient. Beyond unconstrained text-to-image generation, these models now support inpainting and image-to-image translation, which are particularly useful when a practitioner wants to keep some parts of an image fixed while changing others, transfer a new style onto an existing scene, or create realistic data augmentation for downstream computer vision tasks.

Keras Hub recently introduced first-class support for Stable Diffusion 3 through modular components such as the SD3 backbone, text preprocessing utilities, and task-specific wrappers for inpainting, image-to-image, and text-to-image pipelines. Instead of treating these models as black-box web services, a researcher can now script and inspect the full pipeline directly inside a Jupyter notebook. This assignment leverages those components to build an end-to-end experimental notebook and to analyze the behavior of SD3 under different configurations from a computer vision perspective.

Fig. 1. Input astronaut image used for all inpainting experiments. A binary mask preserved the astronaut foreground while enabling background replacement.

The goals of this report are threefold. First, to implement and document an end-to-end SD3 inpainting pipeline using the official Keras Hub API. Second, to investigate how sensitive the inpainted output is to prompt wording, guidance scale, and edit strength, and what that means in practice when a user is trying to "steer" the model. Third, to demonstrate the SD3 image-to-image and text-to-image pipelines and compare their behavior with inpainting on qualitatively rich examples, including an underwater astronaut scene, a cherry-blossom walkway, stylized line-art robots, and portrait stylization.

II. Methodology

All experiments were carried out on a workstation equipped with an NVIDIA RTX 4090 GPU using TensorFlow as the Keras backend. Keras and Keras Hub were installed from their official releases, and the Stable Diffusion 3 (SD3) medium backbone was loaded at a spatial resolution of 512×512

(a) Underwater inpainting result.

(b) Cherry-blossom inpainting result.

Fig. 2. Two inpainting outputs generated from the same masked astronaut input by varying only the prompt. **Underwater prompt:** "An astronaut drifting underwater with glowing jellyfish, bioluminescent trails, and deep blue cinematic lighting." **Cherry-blossom prompt:** "An astronaut on a serene garden path under blooming ivory cherry blossoms with warm sunlight and cinematic depth of field." Stable Diffusion 3 regenerates the background differently for each prompt while preserving the astronaut's structure.

using half-precision (float16) weights for improved memory efficiency. The same backbone served as the foundation for all inpainting, text-to-image, and image-to-image tasks described in this report.

A. Code Availability

The complete implementation used for this assignment, including all inpainting, text-to-image, and image-to-image experiments, is publicly accessible through two Colab notebooks.

The notebook containing the full *inpainting pipeline* experiments, including prompt flexibility analysis, is available here: **Click to open Inpainting Code**.

The notebook for Stable Diffusion 3 *image-to-image* and *text-to-image* generation can be accessed here: **Click to open Image-to-Image & Text-to-Image Code**.

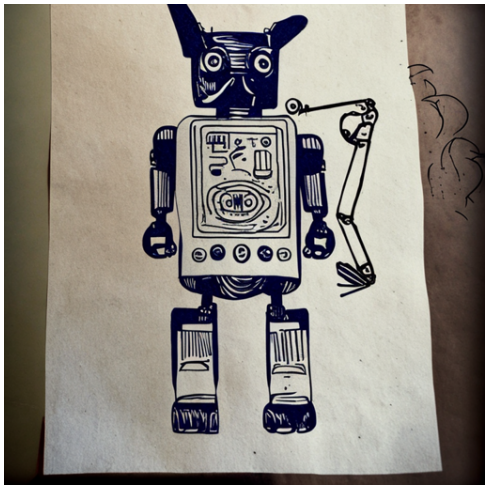
B. Inpainting Pipeline

The inpainting experiments were implemented using the `Stable Diffusion3 Backbone`, `Stable Diffusion3 TextToImagePreprocessor`, and `Stable Diffusion3 Inpaint` modules provided by Keras Hub. Several input images were tested, including a photograph of an astronaut standing on a tree-lined path shown in figure 1. Masks for these images were generated using Google's Gemini 2.5 Image Flash, after which binary masks were manually refined so that the astronaut remained fixed (mask value zero) while the background could be freely altered (mask value one). Each image was resized to 512×512, converted to `float32`, and normalized to the interval $[-1, 1]$. Masks were likewise resized and thresholded into Boolean arrays.

To explore how strongly prompts influence the generated background, the astronaut image was inpainted repeatedly under a wide range of prompts. Each run supplied a descriptive natural-language prompt to guide the model toward specific aesthetics or environmental changes. The SD3 inpaint function was then executed with a configuration dictionary containing the normalized image, the binary mask, and the prompt. Sampling parameters included the number of diffusion steps, the guidance scale that determines prompt adherence, and the strength parameter that controls how far the edited region deviates from the original pixels. A grid search was performed using guidance scales $\{5.0, 7.5, 10.0\}$ and strengths $\{0.4, 0.6, 0.8\}$. A high diffusion step count of 2000 was chosen for certain runs to prioritize image fidelity over runtime. Model outputs were returned as unsigned 8-bit images, stored, and visualized for analysis.

C. Prompt Flexibility Study

Figure 2 shows that the prompt flexibility investigation kept the astronaut photograph and its mask fixed while varying only the textual prompt and sampling parameters for cherry-blossom and underwater scenario. Prompts ranged from an underwater jellyfish environment to a cherry-blossom garden walkway and other stylistic changes. For each prompt, the same grid of guidance scales and strength values was applied. The resulting images were assessed qualitatively for structural preservation, stylistic transformation, and semantic alignment



(a) Original robot line-art.

(b) Inpainted robotic bulldog.

Fig. 3. Original robot sketch (left) and the inpainted robotic French bulldog output (right) generated by Stable Diffusion 3.

with the described scene. By controlling all factors except text conditioning, the study isolated how sensitively SD3 responds to variations in prompt design.

D. Text-to-Image Generation

Text-to-image generation was performed using the `StableDiffusion3TextToImage` wrapper built on the same SD3 backbone. A diverse set of rich aesthetic prompts was constructed to showcase the full generative range of the model shown in figure 5. These included scenes such as a glowing moonlit forest sanctuary, a neon samurai alley in a cyberpunk city, and a floating library suspended above the clouds. For each prompt, the model generated one or more images using moderate guidance scales and 40–60 diffusion steps. The outputs were subsequently inspected for prompt fidelity, composition, and overall visual quality.

E. Image-to-Image Translation

The image-to-image pipeline was implemented using the `StableDiffusion3ImageToImage` wrapper. A real portrait photograph of the author was selected as the reference image shown in 6a, resized to 512×512, and normalized to the $[-1, 1]$ range. This reference was then stylized under several prompts while retaining the subject's identity and pose. Example prompts included a rainy cyberpunk night with holographic neon reflections, a soft watercolor illustration with pastel tones and paper textures, and a winter portrait featuring falling snow and cool-blue ambience. The image-to-image model was invoked with the reference image and the chosen prompts, using sampling parameters similar to those in the inpainting setup. A strength value of approximately 0.6–0.7 was found to provide a strong stylistic transformation without distorting the subject's facial geometry.

III. Results

A. Inpainting Quality

The inpainting pipeline successfully replaced the original tree-lined background behind the astronaut with rich alternative environments while preserving the astronaut's pose, suit geometry, and overall lighting coherence. For the underwater jellyfish prompt, the model generated visually convincing water caustics, volumetric rays streaming from the surface, and large bioluminescent jellyfish surrounding the astronaut, as shown in Fig. 2a. The transition between the preserved foreground and the synthesized background remained smooth, particularly when the binary mask was carefully drawn to avoid intersecting the astronaut's silhouette.

For the cherry-blossom prompt, the same astronaut was integrated into a serene walkway framed by blooming trees and drifting petals (Figs. 1 and 2b). The model maintained the approximate horizon, camera perspective, and scene geometry from the original photograph while regenerating the entire environment according to the new textual description. Increasing the classifier-free guidance scale generally improved semantic alignment with the prompt, whereas higher strength values produced more dramatic background changes at the cost of slight deviations in lighting consistency. We also observed that using a higher number of diffusion steps significantly increased computation time but consistently yielded sharper and more detailed results.

A similar behavior is evident in the line-art experiment, where Stable Diffusion 3 transformed a retro robot sketch into a stylized robotic French bulldog while preserving the parchment texture and ink-drawing aesthetic. The original and inpainted outputs are shown side-by-side in Fig. 3, illustrating SD3's flexibility even on non-photographic, hand-drawn inputs.

B. Prompt Flexibility

In figure 2, the prompt flexibility study revealed that SD3 is highly responsive to coarse semantic changes in the text while maintaining stable foreground structure under inpainting. When the prompt switched from an underwater scenario to a garden or other environment, the background appearance changed dramatically in style, color palette, and local details. At moderate guidance scales, the model balanced adherence to

Fig. 4. Nine inpainting results generated by varying strength (rows) and guidance scale (columns) for the same masked astronaut input. Strength controls the amount of noise injected into the masked region, while guidance controls prompt adherence. All images were generated with 2000 diffusion steps.

(a) Cyberpunk Samurai (b) Steampunk Observatory (c) Desert Nomad

Prompts: (a) A neon-lit samurai alley in a rainy cyberpunk city. (b) A grand steampunk observatory with brass gears and glowing light. (c) A desert nomad under the Milky Way, warm lantern light and dunes."

Fig. 5. Text-to-image outputs from Stable Diffusion 3 for three cinematic prompts, spanning diverse styles from cyberpunk to steampunk and desert night scenes.

the prompt with preservation of global composition. Extremely high guidance sometimes produced oversaturated colors or repeated patterns, whereas very low guidance resulted in weaker alignment with the described scene. Overall, the same masked astronaut could be convincingly integrated into multiple imaginative contexts through prompt changes alone.

The line-art robot example provided an additional view of flexibility. Starting from a retro robot drawing on parchment, an inpaint prompt describing a droopy French bulldog with mechanical elements yielded an output that preserved the ink-on-paper aesthetic while transforming the subject into a robotic dog illustration (Figs. 3a and 3b). This demonstrated that SD3 can adapt not only photographic scenes but also stylized vector-like inputs, as long as the prompt carefully describes both content and style.

C. Text-to-Image Synthesis

In figure 5, the text-to-image experiments showed that the SD3 Text-to-Image pipeline can generate visually rich and coherent scenes directly from descriptive prompts. For the moonlit forest sanctuary, the model produced a glowing moon, dense trees, mist layers, and fireflies consistent with the narrative description. The cyberpunk samurai alley prompt resulted in towering neon signs, rain-soaked streets, and a strong color contrast between magenta and cyan, forming an image reminiscent of concept art. Fantasy prompts such as a floating library above the clouds or a luminous whale gliding through a starry sky were also rendered with high detail and consistent lighting.

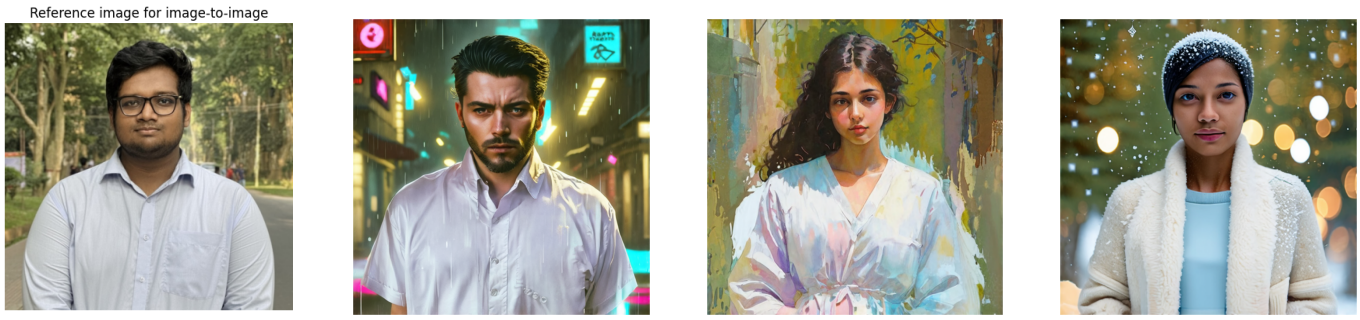
Although different random seeds produced different compositions, the overall semantic content remained stable for each prompt. The model tended to exaggerate cinematic lighting and color grading, which is desirable for aesthetic outputs but may need to be moderated for more realistic applications.

D. Image-to-Image Portrait Stylization

In figure 6, the SD3 Image-to-Image model maintained the identity and pose of the portrait subject while applying substantial stylistic transformations. Under the cyberpunk prompt, the background transformed into a neon-lit city ambience, with colored light reflections appearing on the glasses and face while the overall geometry remained faithful to the original photograph. The watercolor prompt produced softer edges, pastel tones, and a subtle paper texture overlay, giving the impression of a hand-painted illustration. The winter scene prompt added falling snowflakes, cooler color temperature, and background bokeh, again without destroying the subject's facial features.

IV. Discussion

Across all experiments, Stable Diffusion 3 demonstrated strong generative capabilities, but several practical limitations



(a) Reference Image (b) Cyberpunk Style (c) Watercolor Style (d) Winter Scene

Fig. 6. Image-to-image stylization using Stable Diffusion 3. The model retains the subject's identity while applying strong stylistic transformations: (b) cyberpunk neon rain, (c) watercolor painting, and (d) snowy winter aesthetic.

became clear during evaluation. In the inpainting experiments, higher diffusion step counts consistently produced sharper and more coherent results; however, this improvement came with a substantial computational trade-off. Using 2000 diffusion steps on an RTX 4090 required approximately 8 minutes and 24 seconds per image, compared to roughly 3 minutes for 500 steps. Higher-resolution inputs could likely further improve fidelity, but the GPU memory footprint already reached nearly 21 GB during 512×512 inpainting, preventing exploration of larger resolutions within the available hardware constraints.

For text-to-image generation, SD3 performed impressively well. The model produced rich, aesthetic scenes with strong adherence to the descriptive prompts. Although even more refined prompts might yield higher-quality images, the results obtained from the prompts used in this study clearly show SD3's ability to synthesize diverse cinematic environments across cyberpunk, desert, and steampunk themes.

In image-to-image translation, SD3 successfully applied global stylistic transformations such as cyberpunk neon lighting, watercolor rendering, and winter atmospheres. However, when using the author's portrait as the reference image, the model did not fully preserve the subject's facial identity across outputs. While the backgrounds and stylistic elements aligned strongly with the prompts, facial consistency remained a challenge—a known limitation in diffusion-based I2I pipelines unless identity-preserving modules are added.

Overall, the study demonstrates that Stable Diffusion 3 offers high flexibility across inpainting, text-to-image, and image-to-image tasks, but careful tuning of diffusion steps, prompt quality, and hardware considerations is essential for achieving optimal performance.

V. Conclusion

This study demonstrated the inpainting, text-to-image, and image-to-image capabilities of Stable Diffusion 3 using Keras Hub. While the model produced high-quality and visually appealing results, several limitations were observed. Higher diffusion steps improved image fidelity but increased generation time significantly, reaching more than eight minutes per image at 2000 steps. GPU memory usage also reached around 21 GB, preventing experiments with higher resolutions. In image-to-image tasks, strong stylistic prompts sometimes altered facial identity despite preserving pose and background.

Future work includes testing more memory-efficient sampling methods, exploring identity-preserving image-to-image models, and evaluating higher-resolution SD3 variants when hardware allows. Overall, SD3 performs well across tasks but still requires careful parameter tuning for optimal results.