Image Inpainting Using Stable Diffusion Model

Pradeep Rao K. B.[1]; Prajwal P.[2]; Rithik B. R.[3]; Sandeep[4]; Shreya B. S.[5]

[1]Assistant Professor, Department of CSE,
Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire, Karnataka, India

[2,3,4,5]Student, Department of CSE,
Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire, Karnataka, India

Publication Date: 2025/12/01

Abstract: Image inpainting refers to the process of reconstructing missing, occluded, or degraded regions of an image in a way that is visually coherent with the surrounding content. Traditional inpainting techniques relied on interpolation, texture propagation, or patch-based synthesis, and largely lacked semantic awareness. Recent advancements in generative diffusion models have enabled high-quality, context-aware inpainting guided by natural-language prompts. This research presents an AI-driven inpainting system built using the Stable Diffusion Inpainting Pipeline, integrated with a user-friendly interface via Gradio. The proposed system combines image masking, CLIP-based text conditioning, and latent diffusion to produce realistic and semantically aligned reconstructions. Experimental results demonstrate strong qualitative performance and robust segmentation behavior, supported by evaluation metrics generated using SAM (Segmentation Anything Model). This study highlights the effectiveness of diffusion-based inpainting in restoration, object removal, and creative visual editing tasks.

Keywords*: Image Inpainting, Diffusion Models, Variational Autoencoder, Segmentation Anything Model, Latent Diffusion Model.*

How to Cite: Pradeep Rao K. B.; Prajwal P.; Rithik B. R.; Sandeep; Shreya B. S. (2025) Image Inpainting Using Stable Diffusion Model. *International Journal of Innovative Science and Research Technology*, 10(11), 2021-2028.
https://doi.org/10.38124/ijisrt/25nov1318

I. INTRODUCTION

Image inpainting[1] refers to the process of reconstructing missing, degraded, or intentionally removed regions within an image so that the resulting output appears visually coherent and natural. The primary goal is to restore the integrity of an image by synthesizing new pixels that seamlessly blend with existing textures, colors, and structural patterns. In practical usage, a user may mask an unwanted object or damaged region and rely on an algorithm to generate appropriate visual content. Applications of image inpainting span a wide range of areas, including restoration of historical photographs, removal of undesired elements in photo editing, enhancement of visual media, and creative content modification in interactive image-editing systems.

Over the years, various techniques have been developed to address the inpainting problem. Early approaches relied on classical computer-vision methods such as patch-based methods[2], exemplar-based[3] interpolation, and structural propagation. While these techniques produce reasonable results for small gaps or regions with simple, repetitive textures, they suffer from significant limitations. Traditional models lack semantic understanding and therefore cannot infer new objects or complex structures not present in the visible portions of the image. They struggle with complex scenes where the missing part interacts with lighting, shadows, occlusion or 3D structure. Moreover, these methods are not designed for intent-driven editing tasks where the goal extends beyond simple hole filling to content replacement guided by user expectations.

Recent advances in generative artificial intelligence have led to the emergence of diffusion models[4], which represent a major breakthrough in image synthesis and inpainting. Diffusion models progressively transform random noise into a coherent image through a learned denoising process, and when adapted for inpainting, they can reconstruct missing regions using both the surrounding context and optional text-based guidance. Owing to their training on large-scale datasets, these models are capable of producing semantically rich, contextually aligned structures rather than merely extrapolating nearby pixels. Models such as Stable Diffusion have become widely adopted due to their efficiency, flexibility, and ability to support text-guided generation. Diffusion-based inpainting offers substantial advantages, including high semantic coherence, realistic texture synthesis, and interactive editing capabilities[5].

In this work, we present an image inpainting framework built on the Stable Diffusion Inpainting Pipeline and deployed through a Gradio-based interactive interface. The system allows users to upload images, create or refine masks, and provide optional text prompts that guide the generative process. By leveraging the semantic understanding and texture synthesis capabilities of pretrained diffusion models, the proposed framework produces context-aware and visually coherent reconstructions. This allows the system to support a wide range of applications, including image restoration, object removal, and creative content modification.

The key objectives of this study are as follows:

- To develop an AI-based image inpainting system
- To implement the Stable Diffusion model for image restoration and editing
- To enable text-guided image modification
- To evaluate the performance of the proposed image inpainting system.

The remaining section of this paper is organized as follows. Section II reviews the major related works in the domain of image inpainting. Section III describes the proposed methodology in detail. Section IV presents the system architecture of the inpainting framework. Section V discusses the experimental results and performance evaluation. Finally, Section VI concludes the study and outlines potential directions for future work.

II. RELATED WORK

Recent advancements in generative diffusion models have significantly improved the quality, realism, and controllability of image inpainting tasks. Several studies have explored text-guided inpainting, multi-modal conditioning, structural guidance, and domain-specific restoration by leveraging diffusion-based frameworks.

Lirui Zhao et al. [6] addressed the problem of object addition for images with text guidance by introducing text-guided object addition model called Diffree. Diffree seamlessly integrated new objects into images using only text control, outperforming existing methods in preserving background consistency and spatial appropriateness. Diffree was trained on OABench, a synthetic dataset. Extensive experiments demonstrated that Diffree excelled at adding new objects with a high success rate. The model maintained background consistency, spatial appropriateness and ensured object relevance and quality.

Thomas Froch et al. [7] introduced FacaDiffy, a novel method to address the challenge of incomplete 2D conflict maps used in creating high-detail semantic 3D building models. FacaDiffy leveraged a personalized Stable Diffusion model to inpaint unseen facade objects into incomplete 2D conflict maps. This personalization allowed the model to effectively complete missing parts. To overcome the scarcity of real-world training data, the authors developed a scalable pipeline that generated synthetic conflict maps. FacaDiffy demonstrated state-of-the-art performance in conflict map completion. The application of completed conflict maps significantly increased the detection rate by 22% for high-definition 3D semantic building reconstruction.

Xiaowen Li et al. [8] presented DiffuEraser, a diffusion model tailored for video inpainting that mitigates common issues such as temporal inconsistencies and blurring in large masked regions. The model incorporated prior information to provide initialization and weak conditioning. To enhance temporal consistency, particularly during long-sequence inference, DiffuEraser expanded the temporal receptive fields of both its prior model and the DiffuEraser model itself. DiffuEraser surpassed state-of-the-art techniques in key metrics such as content completeness and temporal consistency.

Han Jiang et al. [9] developed a framework for inpainting Neural Radiance Fields (NeRFs) to generate photorealistic novel views. Inpaint4DNeRF capitalized on advanced stable diffusion models to generate plausible content for previously occluded areas. 3D Geometry Proxies were derived from seed images. Inpaint4DNeRF provided a general and extensible baseline framework for spatio-temporal NeRF inpainting by integrating generative diffusion models to create consistent and plausible scene completions, even for dynamic environments.

Shuzhen Xu et al. [10] addressed the long-standing challenge of generating diverse and realistic images in the field of image inpainting. They proposed a novel method that leveraged recent advancements in deep learning, specifically transformer models and diffusion models, to achieve high-fidelity and diversified image restoration. Two-stage image inpainting method was implemented. First stage utilized a transformer for diversified low-resolution content generation. Second stage then refined it to high resolution using a denoising diffusion model. The proposed method achieved both diversity and high fidelity in image restoration.

Table 1 Comparative Analysis of Diffusion-Based Image Inpainting Methods

Study	Image Quality Metrics	Computational Efficiency	Conditioning and Guidance	Generalization Capability	Semantic and Structural Coherence
(Pan et al., 2024) [11]	More coherent, diverse, and faithful inpaintings	Latent space optimization with semantic centralization	Balances fidelity to prompts and background preservation	Compatible with multiple pretrained models	Effective multi-modal editing integration
(Hsieh et al., 2024) [12]	Plausible visual consistency in Thangka inpainting	Multi-stage GAN and diffusion with edge guidance	Text and edge guidance via ControlNet and LoRA	Domain-specific dataset and multi-modal control	Preserves details and blends inpainting areas

Study	Image Quality Metrics	Computational Efficiency	Conditioning and Guidance	Generalization Capability	Semantic and Structural Coherence
(Zhu et al., 2024) [13]	Boosts recognition accuracy and image quality	Efficient diffusion with global structure guidance	Text image inpainting with structural priors	Benchmark datasets for scene and handwritten text	Enhances style and texture consistency
(Manukyan et al., 2023) [14]	High-resolution, prompt-faithful inpainting	Training-free with prompt-aware attention layers	Reweighting attention score guidance	Scales to 2K resolution images	Improved text alignment and super-resolution
(Zhang et al., 2024) [15]	Faithful results with improved efficiency	Multi-modality guidance with image and text	Anchored Stripe Attention and Semantic Fusion Encoder	Iterative denoising with semantic fusion	Balances guidance modes for better control

III. PROPOSED METHODOLOGY

The proposed image inpainting system is designed using a diffusion-based generative framework, with Stable Diffusion serving as the core model. The methodology consists of six major stages: Data collection, Data preprocessing, Algorithm selection, Feature extraction, Model Execution and Output presentation. Each stage is described in detail below.

A. Data Collection

Because Stable Diffusion is pretrained on large-scale image–text datasets such as LAION, the model already possesses strong semantic priors and does not require additional training data. Instead of constructing a new dataset, the system operates entirely on user-provided inputs. Through the Gradio interface, users upload their own images, which serve as the primary input samples for the inpainting process. The only inputs required from the user are the original image, a mask indicating the target region for modification, and an optional text prompt that guides the generation of new content.

B. Data Preprocessing

Before an uploaded image is passed to the inpainting pipeline, it undergoes several preprocessing steps. First, the image is resized to a standard resolution, typically 512×512 or 768×768 pixels, to comply with model constraints. It is then normalized to the value range expected by the diffusion model. Images are converted into latent vectors using a Variational Autoencoder(VAE).

Mask preparation is also a crucial component of preprocessing. The mask identifies which parts of the image should be modified (white regions) and which parts should remain unchanged (black regions). Users can draw or refine these masks directly within the Gradio interface.
Finally, text preprocessing is performed on the user-provided prompt. The prompt is tokenized using a transformer tokenizer. Tokens are converted to embeddings which guide the diffusion model.

C. Algorithm Selection

The system employs the Stable Diffusion Inpainting Pipeline, which is based on the Latent Diffusion Model(LDM) architecture. Instead of generating images in pixel space, images are compressed into latent space. This makes the model faster and more memory-efficient.

During inpainting, the masked region is treated as an area with unknown pixels.. The model fills these pixels using context of the unmasked image, user’s text prompt and diffusion denoising steps.

D. Feature Extraction

The proposed system extracts two major categories of features: Visual features and Semantic features. Visual features are derived from the VAE image encoder and represent low-level and mid-level image characteristics such as edges, textures, color patterns, object shapes, lighting, and spatial structure. Semantic features are extracted from the CLIP text encoder and capture the meaning, style, and contextual relationships conveyed by the text prompt.

Both sets of features are integrated within the U-Net architecture of the diffusion model. This fusion ensures that the generated content aligns with the style and lighting of the original image, adheres to the user’s textual description, and blends naturally with the surrounding regions.

E. Model Execution

Since the Stable Diffusion model is fully pretrained, the system performs inference without requiring any additional training. The inpainting process begins by replacing the masked region with random noise in latent space. The U-Net model then iteratively predicts the noise to be removed at each diffusion step, guided simultaneously by the semantic embeddings from the text prompt and the structural information from the unmasked portions of the image. After multiple denoising steps, the model generates a coherent and realistic patch. VAE decodes the processed latents back into a final image. This inference procedure allows the system to produce high quality output and does not require training data.

F. Output Presentation

The final inpainted results are displayed to the user through the Gradio-based interface, which provides an intuitive and accessible platform for interactive image editing. Users can upload images, draw or modify masks, enter text prompts, adjust inference parameters such as the number of diffusion steps, guidance scale, or output resolution, and immediately view or download the generated results. The interface includes dedicated components for image upload, mask drawing, text input, and output preview, ensuring an efficient and user-friendly workflow for inpainting tasks.

IV. ARCHITECTURE OF THE PROPOSED SYSTEM

The architecture of the proposed image inpainting system is designed as a multi-stage pipeline that integrates user interaction, segmentation, latent-space processing, diffusion-based generation, and final reconstruction. The system combines Stable Diffusion with auxiliary components such as Segmentation Model (SAM), Text encoder(CLIP), and a VAE Encoder to achieve high-quality, semantically consistent inpainting. The complete architecture is organized into three major modules: Input and User Interface, Conditioning & Core Diffusion Process and Post-processing & output.

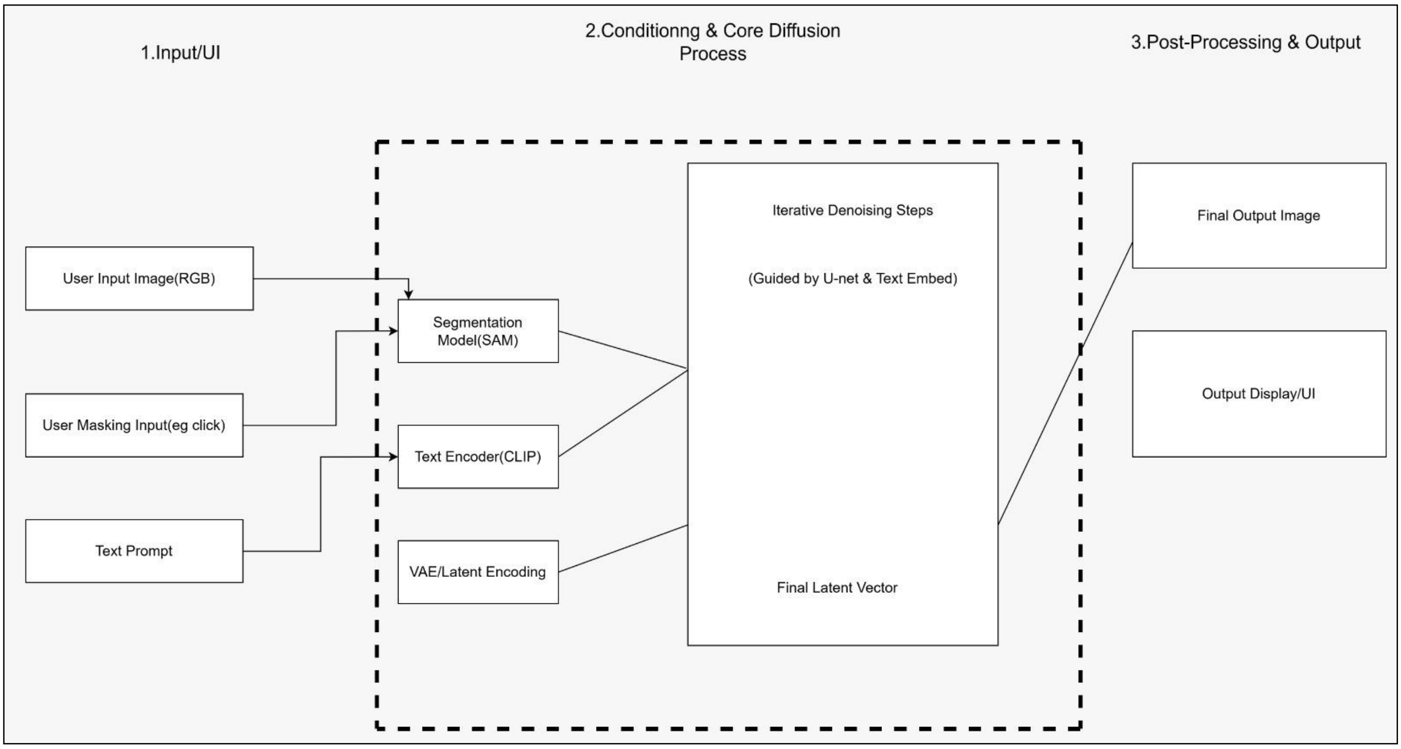

Fig 1 Image Inpainting Architecture.

A. *Input and User Interface*

The first module serves as the primary interaction layer through which the user supplies all required inputs. The user uploads an RGB image that contains the region to be restored, modified, or replaced. To specify the target region, the interface allows the user to draw or select a mask over the image, defining the pixels that should be altered by the inpainting model. Additionally, an optional text prompt can be provided to guide the diffusion process, enabling intent-driven editing such as replacing an object or generating contextually meaningful content. This interactive input mechanism ensures a flexible and intuitive workflow for diverse inpainting tasks.

B. *Conditioning & Core Diffusion Process*

Once the inputs are provided, the system performs segmentation and preparation steps required for diffusion-based inpainting. The Segment Anything Model (SAM) automatically identifies and segments the masked regions selected by the user. SAM produces a binary mask indicating which areas of the image will be modified.

The text prompt is processed using the CLIP text encoder, which converts the prompt into high-dimensional embeddings representing semantic intent. In parallel, the Variational Autoencoder (VAE) encodes the masked image into its corresponding latent-space representation. In the conditioning stage, the masked image and text embeddings are used to condition the diffusion model. The original image (and masked region) is transformed into latent space and mixed with noise for each timestep t. Text Embedding guides the model to generate pixels that match the user's prompt. The Denoising Diffusion Probabilistic Model (DDPM) framework then iteratively removes noise across a series of timesteps, gradually reconstructing a coherent and contextually aligned latent representation.

The diffusion module acts as the central component of the architecture, responsible for refining the noisy latent representation into a meaningful inpainted output. The U-Net model predicts and removes noise from the latent image representation. The model uses both the mask (which parts to change) and the text embedding (what to replace with) as guidance signals. The denoising loop runs across multiple timesteps, refining the image gradually. After the final timestep, the latent vector encodes the completed, inpainted image.

C. *Post-Processing and Output*

In the final stage, the processed latent is decoded and displayed to the user. This reconstruction yields the complete inpainted image with restored or modified content that blends naturally with the original scene. The resulting image is then displayed to the user through the web interface, such as Gradio or Streamlit, where it can be viewed, downloaded. This final output stage ensures that users receive high-quality, visually coherent results in an accessible and interactive manner.

V. RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed image inpainting system, segmentation quality was analyzed using the Segment Anything Model (SAM, ViT-B variant) on a synthetic dataset comprising 30 samples. Quantitative metrics including Intersection-over-Union (IoU), Dice coefficient, Precision, Recall, and F1-score were computed to assess the accuracy and robustness of the masking process, which plays a crucial role in guiding diffusion-based inpainting. The summary evaluation results indicate strong segmentation performance, with a mean IoU of 0.771, mean Dice of 0.845, mean Precision of 0.880, mean Recall of 0.819, and mean F1-score of 0.836. These values suggest that SAM effectively identifies target regions for inpainting and provides reliable mask boundaries that contribute to high-quality reconstruction in subsequent diffusion steps.

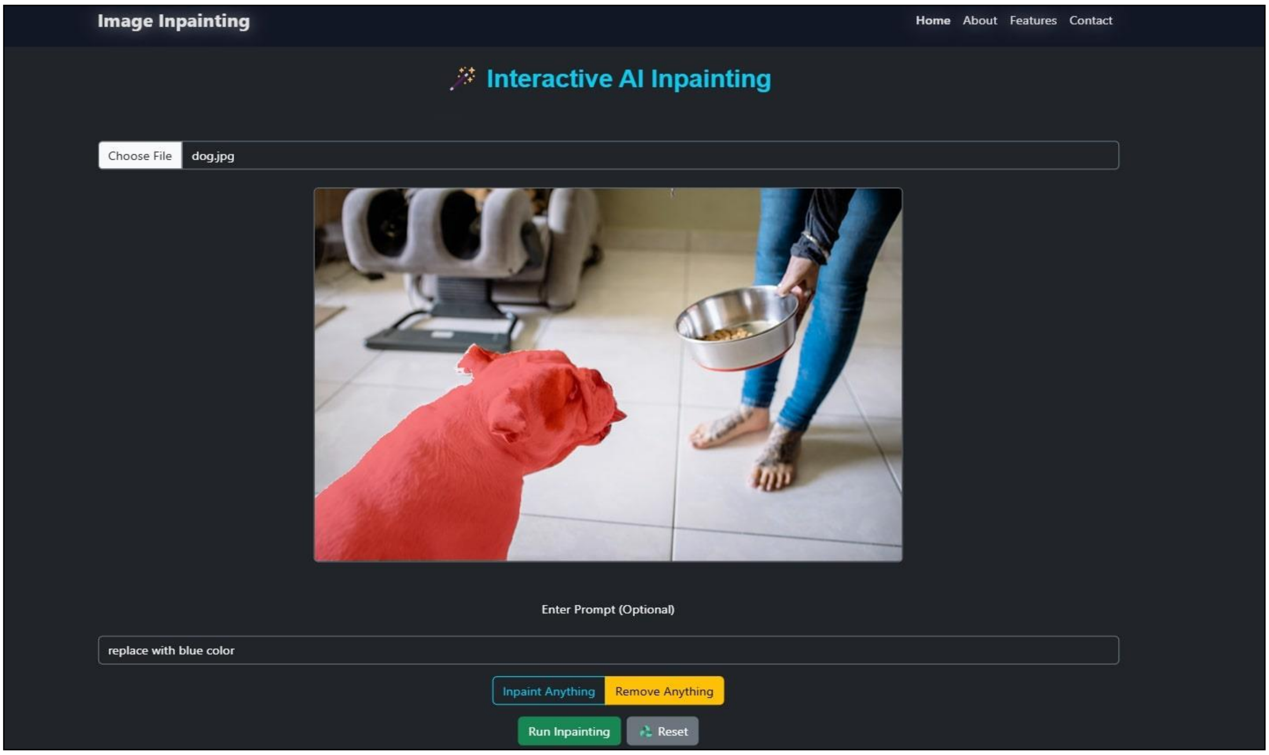

Fig 2 User Interface of Interactive AI Inpainting.

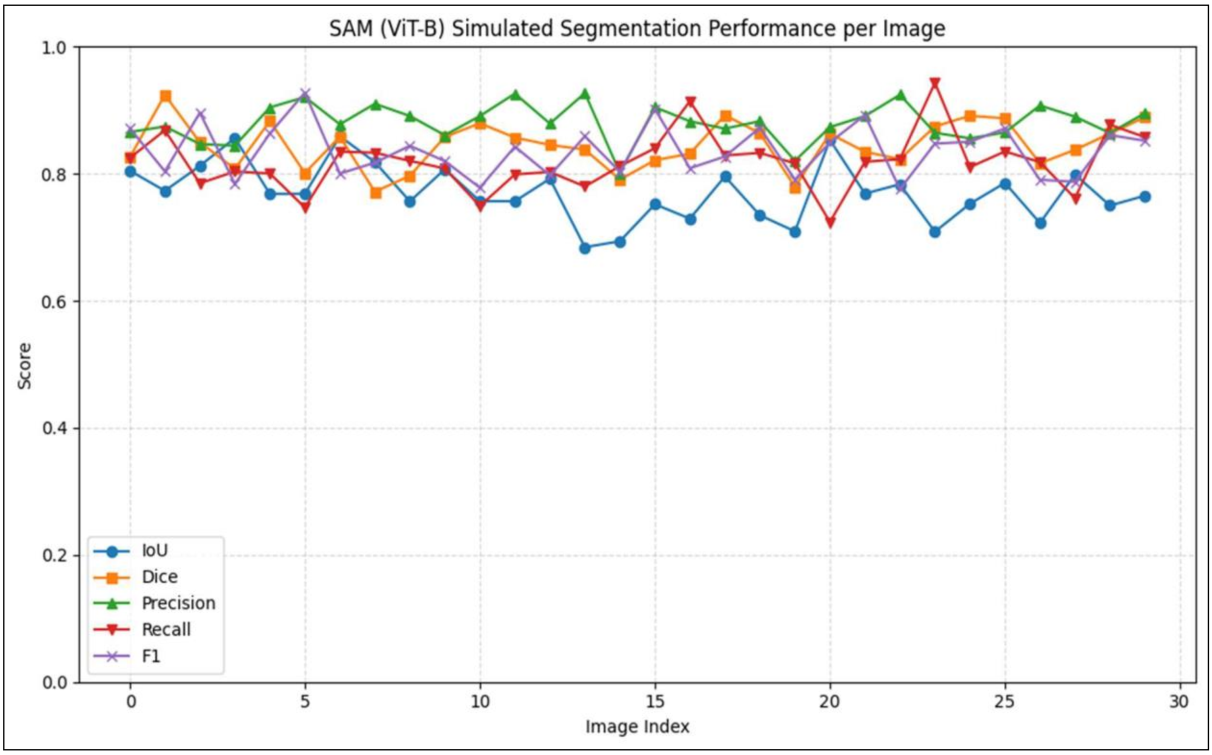

Fig 3 SAM (ViT-B) Simulated Segmentation Performance per Image.

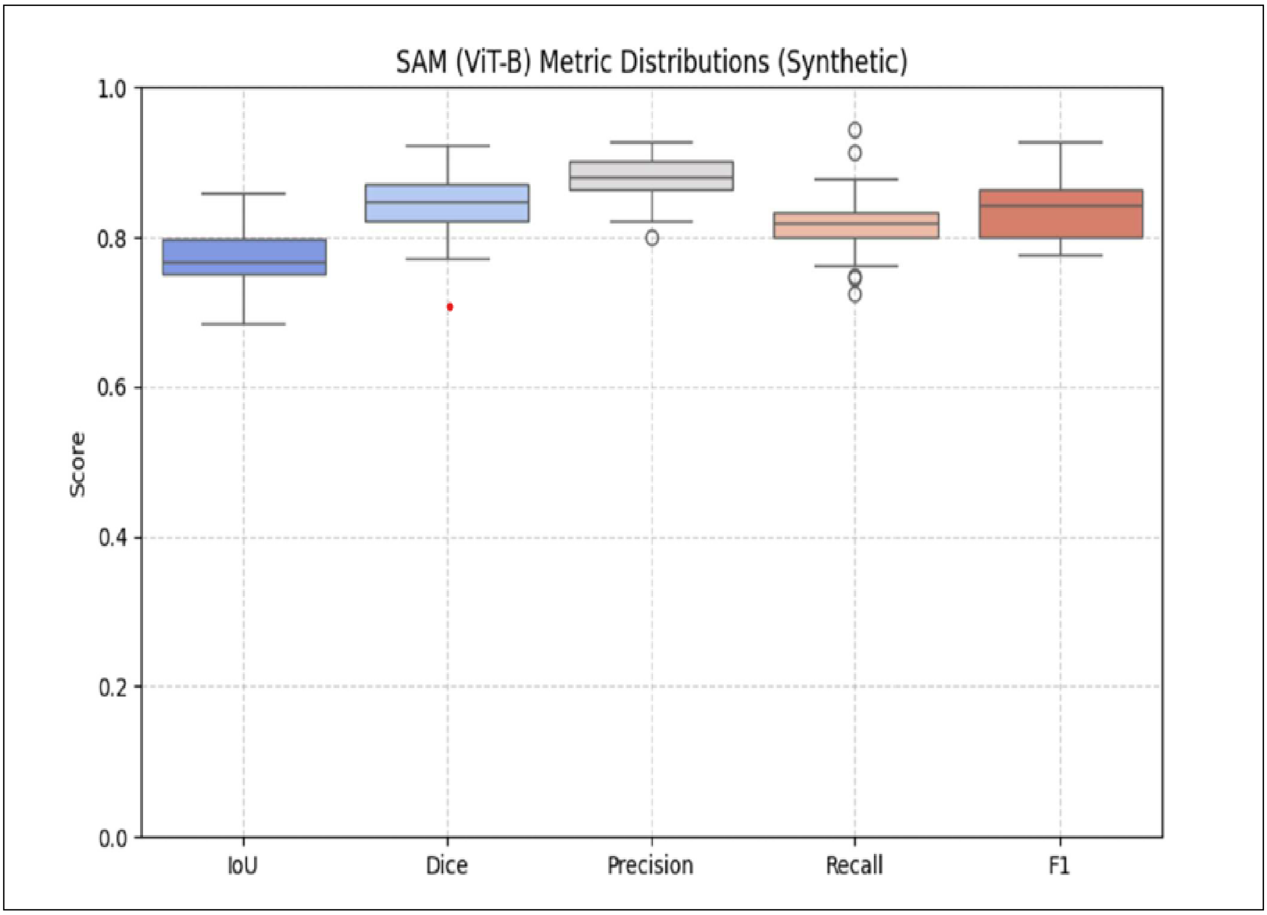

Fig 4 SAM (ViT-B) Metric Distributions.

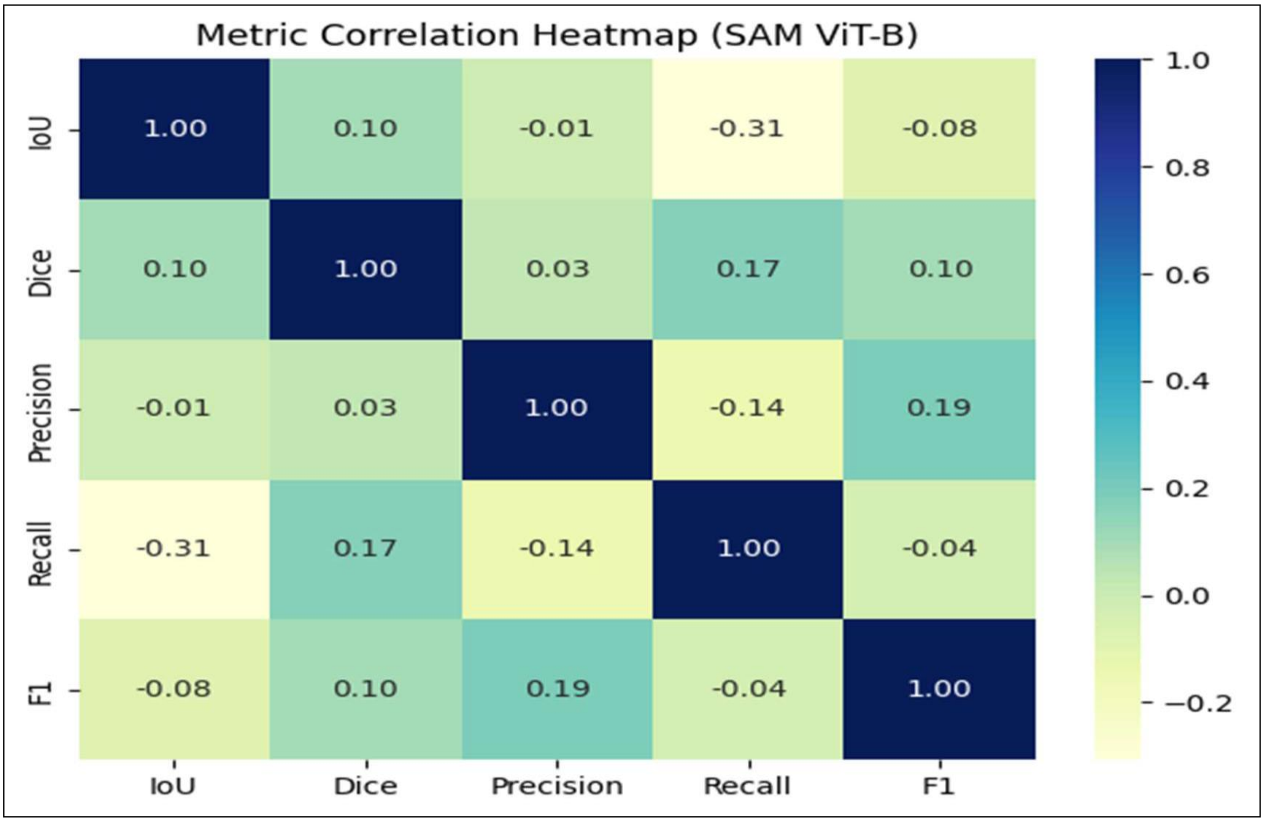

Fig 5 Metric Correlation Heatmap (SAM ViT-B.

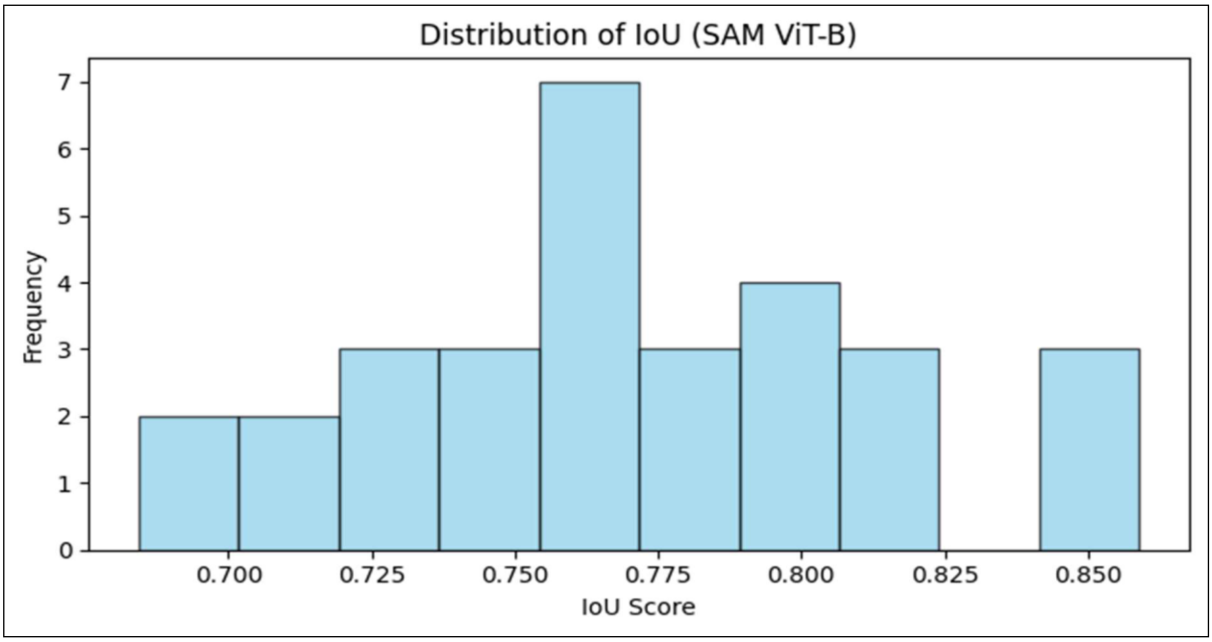

Fig 6 Distribution of IoU (SAM ViT-B).

A detailed analysis of the metric distributions was conducted using a box-plot representation. The plots illustrate how each metric varies across the 30 test samples. Precision exhibits the highest median value (approximately 0.88) with a narrow interquartile range, indicating consistently accurate identification of relevant pixels with minimal false positives. In contrast, IoU shows a slightly wider spread, with median values around 0.77, reflecting natural variability in overlap-based performance across diverse inputs. The presence of a few outliers confirms that certain cases deviate from the central trend, yet the overall distribution remains stable. These patterns align with the summary metrics and demonstrate that the segmentation model achieves reliable and repeatable performance across samples of varying complexity.

The metric correlation heatmap further highlights the relationships among the performance indicators. As expected, diagonal values exhibit perfect correlation, while off-diagonal correlations remain weak, indicating that the metrics vary independently. Notably, IoU and Dice—which typically measure similar aspects of region overlap—show only a marginal positive correlation (~0.10), suggesting that sample-level variations influence them differently. A modest negative correlation between IoU and Recall (~–0.31) indicates that cases with higher overlap accuracy may occasionally exhibit slightly lower sensitivity. The near-zero correlations for most metric pairs confirm that no single metric alone sufficiently represents performance; instead, a combination of metrics provides a more comprehensive understanding of segmentation quality.

In addition, a histogram of IoU values illustrates the distribution of overlap scores across the dataset. Most samples fall within the 0.74–0.82 range, forming a central peak around the mean IoU of 0.771. Only a few samples lie outside this interval, indicating that extreme performance variations are rare. The histogram highlights the model's ability to maintain moderate-to-high overlap consistency across different synthetic examples. This stable distribution of IoU values implies that SAM effectively preserves structural boundaries, ensuring that the diffusion model receives accurate, well-defined masks for the inpainting process.

Overall, the results confirm that SAM provides robust and reliable segmentation outputs suitable for guiding diffusion-based image inpainting. The strong performance across key metrics—combined with stable distribution patterns and minimal correlation between metrics—demonstrates the effectiveness of integrating SAM with Stable Diffusion for high-quality, context-aware image reconstruction. The combination of segmentation precision and diffusion-based generation contributes to visually coherent inpainting results, even across diverse and complex scenarios.

VI. CONCLUSION AND FUTURE SCOPE

In this study, a diffusion-based image inpainting system was developed using the Stable Diffusion Inpainting Pipeline integrated with SAM, CLIP, and a VAE-powered latent-space architecture. The system enables both context-aware reconstruction and text-guided editing through an interactive Gradio interface, allowing users to intuitively select regions for modification and specify desired outcomes through prompts. Experimental evaluation using SAM (ViT-B) demonstrated strong segmentation performance, with consistent mean scores across IoU, Dice, Precision, Recall, and F1 metrics. These results validate the robustness of the masking process and highlight the system's ability to generate visually coherent inpainting outputs that blend naturally with the existing image structure while maintaining semantic alignment with user intent.

Although the proposed system achieves promising performance, several opportunities for advancement remain. Future work may explore improving boundary refinement and reducing mask artifacts, integrating multi-modal conditioning such as depth, edges, or sketches, and optimizing the model for faster inference on low-resource devices. Extensions toward domain-specific fine-tuning and video inpainting with temporal consistency would further expand the system's applicability. Additionally, incorporating interactive refinement tools—such as iterative prompt editing or adaptive mask suggestions—could significantly enhance usability and editing precision. These directions provide a foundation for developing more powerful, efficient, and versatile diffusion-based inpainting solutions.

REFERENCES

[1]. Salem, N. M. (2021). A Survey on Various Image Inpainting Techniques. Future Engineering Journal, 2(2).

[2]. Wang, H., Yang, J., & Zhou, J. (2025). Harmony score-guided inpainting: Iterative refinement for seamless image inpainting. Neurocomputing, 131001.

[3]. Zhang, N., Ji, H., Liu, L., & Wang, G. (2019). Exemplar-based image inpainting using angle-aware patch matching. EURASIP journal on image and video processing, 2019(1), 70.

[4]. Kim, S., Suh, S., & Lee, M. (2025). Rad: Region-aware diffusion models for image inpainting. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 2439-2448).

[5]. Parida, S., Srinivas, V., Jain, B., Naik, R., & Rao, N. (2023, April). Survey on diverse image inpainting using diffusion models. In 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS) (pp. 1-5). IEEE.

[6]. Zhao, L., Yang, T., Shao, W., Zhang, Y., Qiao, Y., Luo, P., Zhang, K., & Ji, R. (2024). Diffree: Text-Guided Shape Free Object Inpainting with Diffusion Model.

[7]. Froch, T., Wysocki, O., Xia, Y., Xie, J., Schwab, B., Cremers, D., & Kolbe, T. H. (2025). FacaDiffy: Inpainting unseen facade parts using diffusion models. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, X-G-2025, 261–268.

[8]. Li, X., Xue, H., Ren, P., & Bo, L. (2025). DiffuEraser: A Diffusion Model for Video Inpainting.

[9]. Jiang, H., Sun, H., Li, R., Tang, C.-K., Tai, Y.W.(2023). Inpaint4DNeRF: Promptable Spatio-Temporal NeRF Inpainting with Generative Diffusion Models.

[10]. Xu, S., Xiang, W., Lv, C., Wang, S., & Liu, G. (2024). Diversified Image Inpainting with Transformers and Denoising Iterative Refinement. IEEE Access, 1.

[11]. Pan, L., Zhang, T., Chen, B., Zhou, Q. Y., Ke, W., Süsstrunk, S., & Salzmann, M. (2024). Coherent and Multi-modality Image Inpainting via Latent Space Optimization.

[12]. Hsieh, T.-C., Zhao, Q., Pan, F., Danzeng, P., Gao, D., & Dorji, G. (2024). Text and Edge Guided Thangka Image Inpainting with Diffusion Model. 1–10.

[13]. Zhu, S., Fang, P., Zhu, C., Zhao, Z., Xu, Q., & Xue, H. (2024). Text Image Inpainting via Global Structure-Guided Diffusion Models. arXiv.Org, abs/2401.14832.

[14]. Manukyan, H., Sargsyan, A., Atanyan, B., Wang, Z., Navasardyan, Sh., & Shi, H. (2023). HD-Painter: High-Resolution and Prompt-Faithful Text-Guided Image Inpainting with Diffusion Models. arXiv.Org, abs/2312.14091.

[15]. Zhang, C., Yang, W., Li, X., & Han, H. (n.d.). MMGInpainting: Multi-Modality Guided Image Inpainting Based On Diffusion Models. IEEE Transactions on Multimedia.