International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

Pictorama: Text based Image Editing using Diffusion Model

Teena Varma[1], Harshali Patil [*2], Kavita Jain[3], Deepali Vora[4], Akash Sawant[5], Vishal Mamluskar[6], Allen Lopes[7], Nesan Selvan[8]

Submitted: 04/02/2024 **Revised**: 12/03/2024 **Accepted**: 18/03/2024

Abstract: This research aims to pioneer image modification through text, integrating natural language descriptions with advanced computer vision and NLP techniques. The primary objective is to bridge human language and image editing, empowering users to convey creative visions effortlessly, revolutionizing the field of image modification. The study employs stable diffusion models, leveraging PyTorch and Python. It builds on prior works like Imagic, LEDITS, and Instructpix2pix, integrating a novel Vector Quantized Diffusion (VQ-Diffusion) model. The model is trained on a dataset of 436 GB containing 3 features an input image, an editing instruction, and an output edited image. Test samples include real images subjected to diverse text prompts for image edits, with disentanglement properties explored. The approach combines text inversion and Box-Constrained Diffusion (BoxDiff) for personalized and conditional image synthesis. The research showcases that stable diffusion models exhibit disentanglement properties, enabling effective modifications without extensive fine-tuning. The introduced BoxDiff and VQ-Diffusion models demonstrate superior performance in spatially constrained and complex scene synthesis, outperforming traditional methods. We are able to observe greater quality in output images with good cohesiveness throughout the image. Runing the model with greater number of steps allows for upheaval in quality. Here we have used 100 steps for greater image quality. The effect of number of steps on the time taken for inference is also studied. Due to the large amount of video memory required for inferencing, we recommend a GPU with >11 GB of video memory. The study adds value by addressing biases, achieving higher speeds, and enhancing image quality, contributing to the evolving landscape of text-to-image synthesis. This research introduces novel approaches in disentanglement, spatially constrained synthesis, and rapid image generation, pushing the boundaries of text-to-image synthesis beyond existing limitations.

Keywords: *Image Processing, Computer vision, Pytorch, Stable Diffusion Models, Python, Machine learning.*

1. Introduction

Image editing through text instruction represents a groundbreaking frontier at the confluence of natural language processing (NLP) and image processing, offering an intuitive means to manipulate digital im-ages. While traditional image editing tools demand a steep learning curve and intricate manual adjust-ments, this innovative paradigm enables users to articulate their editing preferences through plain text instructions. Leveraging advanced NLP models like GPT-3, these directives are translated into precise image edits, democratizing the art of image enhancement and fostering a realm of creative possibilities. However, despite recent advancements, significant gaps persist within the existing literature. Seminal works from 2020 to 2024, including Imagic[1], LEDITS[2], and Instructpix2pix[3], have demonstrated the potential of text-driven image editing. Yet, limitations such as generative biases, scalability con-straints, and complex scene synthesis hinder widespread adoption and optimal performance.

To address these limitations, our research endeavors to pioneer novel approaches in disentanglement, spatially con-strained synthesis, and rapid image generation, pushing the boundaries of text-to-image synthesis be-yond existing constraints. By distilling key insights from recent milestone works and highlighting their limitations, we aim to justify the need for our present study. Specifically, our work seeks to rectify the identified gaps by introducing stable diffusion models, leveraging state-of-the-art techniques like Vec-tor Quantized Diffusion (VQ-Diffusion)[14] and Box-Constrained Diffusion (BoxDiff)[13] to enhance the efficiency, scalability, and quality of text-driven image editing.

2. Literature Survey

In the paper authored by Kawar et al.[1], Imagic is introduced as an innovative approach to image ed-iting using natural language text descriptions. The method utilizes a pre-trained text-to-image diffusion model to achieve realistic image generation. The system consists of an image generator and an image editor, demonstrating superior performance in various editing tasks compared to existing methods such as Text2LIVE, DDIB, and SDEdit, particularly in terms of image quality and fidelity. A conducted user study confirms Imagic's user-friendly nature and its effectiveness in effortlessly producing high-quality images.

[1] *Prof. Computer Eng., Xavier Institute of Engineering, Mumbai, INDIA*
ORCID ID : 0009-0000-4594-9314
[2] *Prof. Computer Eng., Thakur College of Engineering and Technology, Kandivali, INDIA*
ORCID ID : 0000-0003-2052-9940
[3] *Prof. Computer Eng., Xavier Institute of Engineering, Mumbai, INDIA*
ORCID ID : 0009-0007-6103-2515
[4] *Prof. Computer Eng., Symbiosis Institute of Technology, Pune, INDIA*
ORCID ID : 0000-0003-3969-9800
[5] *Computer Eng., Xavier Institute of Engineering, Mumbai, INDIA*
ORCID ID : 0009-0003-6507-9055
[6] *Computer Eng., Xavier Institute of Engineering, Mumbai, INDIA*
ORCID ID : 0009-0009-8992-7008
[7] *Computer Eng., Xavier Institute of Engineering, Mumbai, INDIA*
ORCID ID : 0009-0008-2881-0933
[8] *Computer Eng., Xavier Institute of Engineering, Mumbai, INDIA*
ORCID ID : 0009-0001-3713-9407
* *Corresponding Author Email: harshali.patil9@gmail.com*

Imagic's versatility and potential impact are highlighted through its wide-ranging applications in graphic design, advertising, and social media.

Tsaban et al.[2] present LEDITS, a real image editing method utilizing DDPM inversion and seman-tic guidance, addressing limitations of current techniques. The paper showcases LEDITS' effectiveness in diverse applications like object removal, colorization, and style transfer. Emphasizing the use and importance of diffusion models in image editing and generation, the authors contribute to advancing this field. LEDITS not only overcomes existing limitations but also provides a promising direction for future research. The paper serves as a notable contribution to the realm of image editing techniques, demonstrating practical applications and highlighting the potential of diffusion models in this context.

Brooks, Holynski et al.[3] introduce "Instructpix2pix," a system adept at learning and executing im-age editing instructions. The paper delves into the challenges of existing text-to-image synthesis models, underscoring the importance of methods capable of accurately interpreting and implementing textual instructions for image editing. The authors provide a comprehensive overview of strategies and techniques in text-to-image synthesis, advocating for improved evaluation metrics, enhanced datasets, and advancements in the context of architectural design and model training. They highlight difficulties in generating complex scenes from textual descriptions and underline the reduced work on scaling this method to higher resolutions. In essence, the paper critically examines current strategies, identifies re-search gaps, and provides valuable insights for advancing the field.

Kawar et al.[4] paper introduces Imagic, a text-based real image editing method utilizing pre-trained diffusion models. Imagic effectively applies non-rigid edits to real images based on free-form natural language prompts, showcasing its versatility in complex semantic edits like posture and composition changes. The method aligns text embedding with both inputs given by the user, image and target text, which allows for fine-tuning the diffusion model for image-specific appearance. Despite generative limitations and biases, Imagic operates without additional inputs, emphasizing its potential for versatile and high-quality real image editing. The paper has spurred interest in community pipelines and tools for practical implementation in various real-world applications.

Miyake et al.[5] literature review introduces "Negative-prompt Inversion," a rapid image inversion method for text-guided diffusion models in image editing. Focusing on the challenge of time-consuming and computationally expensive inversion processes, the authors propose a method achieving comparable reconstruction quality solely through forward propagation. Their approach is over 30 times faster than null-text inversion with a resolution of 512 pixels with up to 50 sampling steps. The paper underscores the potential of negative-prompt inversion in accelerating image editing while preserving high-quality outcomes, marking it as a valuable contribution with implications for future research and applications in the realm of image generation and editing.

Alayrac et al.[6] literature review introduces Flamingo, which is a family of Visual Language Model (VLM) focused on rapid adaptation to new tasks with minimal annotated examples. The paper proposes innovative architectures to integrate trained vision and language only models, accommodating interwoven image and text data seamlessly. Flamingo's flexibility allows training on large-scale multimodal web corpora, enabling robust in-context few-shot learning. Thorough evaluations demonstrate its superior performance with respect to various visual tasks, surpassing models which are fine-tuned on more task-specific dataset. The paper's impact extends to community interest in developing pipelines and tools for real-world applications of Flamingo models.

The literature review "Diffusion Models Beat GANs on Image Synthesis" by Dhariwal et.al (2021)[7] demonstrates that models incorporating diffusion can give better image quality compared to modern generative models. The authors show that through architectural improvements, diffusion models can outperform existing generative adversarial networks (GANs) in terms of image sample quality. We can achieve this for image synthesis by implementing a superior architecture and greatly improve sample quality with the help of classifier guidance. The paper highlights that while diffusion models tend to be slower than generative adversarial networks at sampling because of the use of multiple steps for de-noising and forward passes, they give preferred properties including distribution coverage, easy scalability and a stationary training objective. The authors also release their code, contributing to the broader research community

A technique for customizing image generation through text inversion is presented in the literature re-view paper by Rinon Gal et al. (2022)[8]. The authors present a method that flips the text embedding to produce a latent code that can be utilized to create personalized images, allowing text-to-image generation to be personalized. The study presents the method's efficacy in producing customized images for a range of uses, such as image editing and facial synthesis. The authors also present a comparison of their approach's advantages in terms of flexibility and personalization with other text-to-image generation techniques. The paper offers a promising direction for future research in this field and advances the state of modern text-to-image generation techniques.

The literature review "Image Inpainting: A Review" by Elharrouss et al. (2020)[9] provides an over-view of current image inpainting approaches, which have been classified as such: sequential-based, CNN-based, and GAN-based. The limitations and difficulties of conventional picture inpainting meth-ods are discussed in the paper, including the need for better assessment criteria and datasets and the inability to handle large-scale situations. The authors present a com-prehensive review of the state of the modern in image inpainting, facilitating the comparison of methods and datasets used in the field. The review highlights the importance of automatic image inpainting in computer vision and its various ap-plications, such as image restoration, adversarial perturbation, and image compression.

The paper "CogView: Mastering Text-to-Image Generation via Transformers" by Ding et al. (2021)[10] introduces CogView, a 4 billion parameter Transformer with a tokenizer designed to address the challenge of text to image generation in the general domain. The authors propose CogView as a powerful generative model that can understand cross-modal information, and they demonstrate its effectiveness in advancing text-to-image generation. The paper also presents strategies for finetuning various tasks, like super resolution, style learning, text and image ranking, as well as ways to increase the stability of pretraining. CogView has great performance on the dataset of blurred MS COCO, outperforming preceding models and similar work. The authors also release the code for CogView, contributing to the research community.

The literature review paper "Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors" by Gafni et al. (2022)[11] presents a novel text to image method that addresses the limitations of existing techniques by enabling simple control mechanisms, improving tokenization processes, and adapting classifier free guidance for transformer models. The method, called Make-A-Scene, achieves best in class FID and human evaluation results, generating high quality images with a resolution of 512x512 pixels. The authors demonstrate several novel capabilities, such as text editing with anchor scenes, unknown text prompts, and generation of story illustration. Make-A-Scene contributes to the advancement of text-to-image generation techniques and provides a promising direction for future re-search in this field.

This paper “Uncovering the disentanglement capability in text-to-image diffusion models”[12] delves into the disentanglement properties of stable diffusion models in the domain of im-age generation, focusing on the separation of semantic content and styles within the generated images. The prima-ry inquiry is whether these models inherently possess the capability for effective disentanglement, akin to what has been observed in Generative Adversarial Networks (GANs). The study reveals that stable diffusion models exhibit disentanglement when specific modifications are made to the text embed-dings, allowing for the alteration of certain attributes without affecting overall identi-ty. The authors propose an optimization scheme to determine optimal combination weights for text embeddings, facilitating disentangled image modifications. The results of the experiments demonstrate the models' capacity to effectively separate different ideas and characteristics, especially in image editing tasks, without requiring a great deal of fine tuning.

The paper “Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion”[13] intro-duces Box-Constrained Diffusion (BoxDiff), an innovative approach for conditional image synthesis. BoxDiff provides fine control over the spatial characteristics of synthetic objects or situations during the denoising stage in Stable Diffusion models, in contrast to popular approaches that mostly rely on text cues. The suggested spatial constraints—Inner-Box, Outer-Box, and Corner Constraints, among others—offer direction for creating objects that satisfy user-specified requirements. Interestingly, BoxDiff is an effective solution for conditional image synthesis as it doesn't require a lot of matched layout-image data or further model training. The paper contributes by presenting a method for training-free image synthesis, seamlessly integrating spatial constraints into the denoising process, and show-casing the capability to synthesize diverse visual concepts in line with user-specified spatial conditions.

The paper “Vector quantized diffusion model for text-to-image synthesis”[14] introduces a novel approach, termed Vector Quantized Diffusion (VQ-Diffusion), for the generation of images from textual descriptions. By combining a Vector Quantized Variational Autoencoder (VQ-VAE) with a conditional version of the Denoising Diffusion Probabilistic Model (DDPM), the model overcomes biases and error accumulation issues observed in existing methods. Employing a mask-and-replace diffusion strategy, the VQ-Diffusion model demonstrates superior performance in generating images, outperforming traditional autoregressive models and GAN-based techniques, especially in handling complex scenes. Notably, the model achieves a substantial speedup of 15 times compared to conventional autoregressive methods while maintaining high image quality.

The paper "Text to Image Generation with Semantic-Spatial Aware GAN"[15] focuses on the challenging task of text-to-image generation (T2I), specifically addressing limitations in existing methods that often produce images inconsistent with the textual descriptions. These methods commonly use Conditional Generative Adversarial Networks (GANs) and initialize images from noise with sentence embeddings, refining them iteratively with word embeddings. While this approach can generate images that align with the overall textual theme, individual image regions may lack coherence or be unrecognizable, particularly for specific object descriptions. This paper proposes a novel framework called Se-mantic-Spatial Aware GAN (SS-GAN) to address these issues”.

Table 1. Literature Review

Paper	**Year of Publication**	**Method Used**	**Dataset**	**Findings**
[1]	2023	*TEdBench* (Textual Editing Benchmark)	Not disclosed	Semantics of text instruction is given more attention
[2]	2023	StableDiffusion-v-1-5 checkpoint	Used a pretrained model.	DDPM Inversion with SEGA scheme of denoising process.
[3]	2023	conditional diffusion model InstructPix2Pix	Generated dataset using GPT and stable diffusion	Using a sizable dataset, "InstructPix2Pix" trains a model to alter images in accordance with handwritten instructions. Fast and effective outcomes can be obtained without adjusting.
[4]	2023	image editing with diffusion models.	Used a pretrained model.	In order to generate embedded text that functions with both the given image as input and the

				required text, it uses a pretrained image diffusion model, which enables the diffusion model to obtain the image-specific appearance.
[5]	2023	Negative-prompt Inversion	Not disclosed	It discusses a method which can be used to achieve comparable reconstruction of picture exclusively through forward propagation
[6]	2022	Visual Language Model (VLM)	Uses the Multimodal C4 dataset for training its models.	The model surpasses other models trained on thousands of times more task-based data, achieving excellent results on several benchmarks.
[7]	2021	Diffusion models	Not disclosed	In image synthesis challenges, diffusion models outperform current generation models and GANs in terms of picture quality.
[8]	2022	Textual Inversion	LAION-400M dataset	By using only 3-5 images of a user-provided concept to represent it through new "words" in the embedding space of a frozen text-to-image model, the method enables personalized creation in an intuitive way.
[9]	2020	Sequential-based, CNN-based, and GAN-based.	Paris StreetView, Places, depth image dataset, Foreground-aware, Berkeley segmentation, ImageNet, CelebFaces Attributes Dataset (CelebA), Indian Pines, Microsoft COCO val2014 dataset.	This paper reviews current image inpainting methods, categorizing them into sequential, CNN-based, and GAN-based approaches. It explores techniques for various image distortion types and highlights available datasets (filling a research gap). Real-world evaluations compare the three categories' performance across different datasets and distortions. The review aids researchers by summarizing methods, datasets, and evaluation metrics. Its key contribution is classifying methods and providing datasets for researchers to benchmark their work.
[10]	2021	Evidence Lower BOund (ELBO)	Blurry MS COCO dataset	CogView is a transformer with four billion parameters and a VQ-VAE tokenizer. explains how to fine-tune techniques for a range of downstream tasks, including super-resolution, text-image ranking, style learning, and fashion design, as well as how to stabilize pretraining by eliminating NaN losses. CogView achieves the state-of-the-art FID on the hazy MS COCO da-taset by outperforming previous GAN-based models and a recent work of a similar kind, DALL-E.
[11]	2022	Novel approach to text-to-image generation.	CC12m , CC, and subsets of YFCC100m and Redcaps amounting to 35m text-image pairs.	Introduction of a novel text-to-image generation method that incorporates human priors to improve the realism of the generated images.
[12]	2023	stable-diffusion-v1-4	laion dataset	Study explores diffusion models in text-to-image generation, revealing the capacity to modify style without changing semantic

				content in stable models
[13]	2023	Box constrained Diffusions.	paired layout-image data of COCO-Stuff	Layouts are created to make images through text instruction
[14]	2022	VQ-VAE’s encoder and decoder follow the setting of VQGAN	CUB-200, Oxford-102 , and MSCOCO datasets.	The encoder and decoder of VQ-VAE use the GAN loss to produce a more realistic image by leveraging the VQGAN [16] setting.
[15]	2023	Visual Instruction Inversion (VISII)	e Clean-InstructPix2Pix dataset	In this paper, visual prompt-based image editing is introduced, demonstrating competitive results in efficiency and surpassing text-conditioned systems with just one example pair.

3. Dataset Used

For Image Editing Through Text Instruction, we combine cutting edge machine learning techniques and a comprehensive dataset to achieve accurate results. The model used for the image editing is trained using a supervised learning process. A dataset comprising an input image, an edit instruction, and an output image with the anticipated alteration is used to train the image generating and editing model.

3.1. Dataset used: clip-filtered-dataset [18]

A clip-filtered dataset refers to a curated collection of images that have been filtered or labeled based on their alignment with certain textual prompts or queries. The process involves associating images with specific descriptions or keywords provided in natural language.

3.2. Features in the dataset:

Input: an input image
Edit: an editing instruction
Output: an output edited image

The utilization of this dataset facilitates more precise training and contributes to superior end results. However, owing to the substantial size of the dataset, training on hardware with lower computational capabilities may pose challenges. As a recommended strategy, employing cloud computing resources is advised. Cloud computing offers scalable and powerful computing capabilities, allowing for efficient training of models on large datasets without the constraints imposed by lower-end hardware. This ap-proach ensures optimal utilization of computational resources and enhances the training process for the image editing model

4. Technologies Used

4.1. Python

Python serves as a foundational element in a project centered around image modification using text, highlighting its versatility in managing both textual and image data. The project leverages libraries such as PIL (Python Imaging Library) or its successor, Pillow, allowing developers to effortlessly manipulate images by incorporating textual elements. Python's readability and simplicity are instrumental in the implementation of algorithms for various tasks, including overlaying text onto images and adjusting font styles and sizes. This choice of programming language streamlines the development process, making it accessible for tasks involving the integration of textual information into images.

4.2. Pytorch

PyTorch assumes a central role in a state-of-the-art image modification project that seamlessly inte-grates text. Leveraging PyTorch's robust deep learning capabilities, the project employs neural net-works to establish a potent framework for image processing. The flexibility and dynamic computation graph of PyTorch empower developers to efficiently design and train models for diverse tasks, ranging from text recognition and sentiment analysis to context-aware image modification.

The framework's extensive support for GPU acceleration significantly expedites the processing of large image datasets, enhancing overall computational efficiency. Additionally, PyTorch's user-friendly interface and comprehensive documentation simplify the implementation of intricate neural network architectures. This ease of use facilitates the creation of sophisticated models that seamlessly incorpo-rate textual information into image modification processes, pushing the boundaries of creativity and innovation within this dynamic project.

4.3. Stable Diffusion Model

In a project dedicated to image modification using text, the incorporation of a robust diffusion model introduces a sophisticated approach aimed at enhancing the visual appeal and quality of modified images. The stable diffusion model, recognized for its effectiveness in preserving and manipulating intricate details in images, becomes a powerful tool for achieving nuanced modifications. This model allows the project to intricately blend textual elements with images, ensuring a smooth and visually pleasing fusion.

The adaptability of the stable diffusion model to diverse image features enables fine-tuned adjustments, maintaining the integrity of the original content while accommodating text-based modifications. This approach not only ensures a high level of stability in the modification process but also opens avenues for creatively enhancing visual storytelling through the seamless integration of text and images. The stable diffusion model thus plays a crucial role in achieving both precision and artistic expression in the project's image modification endeavors.

4.4. GPT-3

GPT-3 (Generative Pre-trained Transformer 3) introduces a revolutionary dimension to a text-based image modification project by leveraging state-of-the-art natural language processing capabilities. GPT-3 is an advanced language understanding and generation tool that excels in comprehending and generating contextually appropriate textual descriptions for images. This capability enables the auto-matic generation of contextual information or captions that can guide modifications to images.

Furthermore, the integration of GPT-3 with image processing methods allows the model to be trained to recognize and respond to textual cues. This enables the implementation of imaginative and dynamic changes to images. The synergy between GPT-3's language capabilities and image manipulation pro-cesses opens up innovative possibilities for intuitive and context-aware image transformations driven by textual input. This approach pushes the boundaries of interactive and intelligent image editing, offer-ing a new level of creativity and adaptability in the modification proces.

5. Hardware and Software Requirement

- Nvidia CUDA enabled Graphics card with at least 12 GB Video memory.
- 15 GB free space
- Windows 11
- Python
- Pytorch
- Stable Diffusion

6. Proposed Methodology

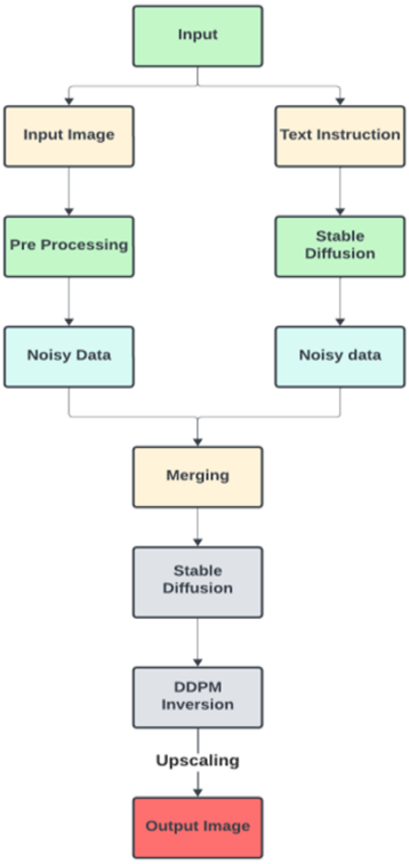


Fig. 1. Proposed Methodology

Components

6.1. Input Image

In this stage, the system takes an image as input, which serves as the base or original image upon which modifications are intended to be applied. This input image becomes the starting point for subsequent processes in the image modification system, where various techniques, models, or methods are employed to achieve the desired alterations or enhancements based on the specified objectives or textual instructions.

6.2. Input Instruction

In this subsequent stage, an additional input is provided in the form of text instructions. These instructions serve as a directive for how the image, after undergoing modification, should be altered or appear. The text instructions play a crucial role in guiding the image modification process, influencing aspects such as content, style, or specific changes that need to be applied to the base image. The integration of textual input adds a layer of context and specificity, allowing the system to interpret and implement the desired modifications in alignment with the provided instructions.

6.3. Pre-processing

In this stage, the image undergoes a preprocessing step to enhance its suitability for computer-based analysis and understanding. The preprocessing involves a series of operations, including but not limited to pixilation, resizing, rotation, and flipping. These operations are designed to optimize the image for subsequent computational processes, making it more amenable to feature extraction, pattern recognition, and other image analysis tasks.

- **Pixilation:** Breaking down the image into smaller, uniform regions, often represented by pixels, to simplify its structure or obscure details.

- **Resizing:** Adjusting the dimensions of the image, either scaling it up or down, to meet specific requirements or to ensure compatibility with the subsequent stages of the image modification process.

- **Rotation and Flipping:** Altering the orientation of the image by rotating it or flipping it horizontally/vertically. These operations can be employed to achieve a desired alignment or orientation for better processing.

Each of these preprocessing steps contributes to creating a more standardized and manageable representation of the image, facilitating a more effective analysis and interpretation by the computer-based system.

6.4. Stable Diffusion

The stable diffusion model describes the random movement of particles, exhibiting stability in their distribution, offering a versatile framework applicable across various scientific contexts.

6.5. Noisy Data

The data generated from stable diffusion is considered noisy, indicating that it lacks the desired structure or format. Noisy data often contains random variations, disturbances, or irregularities that can impede the extraction of meaningful information. In the context of stable diffusion data, additional pro-cessing is required to transform it into a more organized and structured format, aligning it with the de-sired characteristics for image generation.

Further processing methods may involve noise reduction techniques, filtering, or smoothing operations to enhance the signal-to-noise ratio and reveal the underlying patterns within the data. Once the noisy data is appropriately processed, it can be used as a foundation for generating images that align with the intended outcomes of the image modification project.

6.6. DDPM Inversion

In the denoising diffusion probabilistic model, data contaminated with noise undergoes a probabilistic evolution process. This iterative process is designed to refine the data progressively, enhancing the ex-traction of meaningful signals. The model adopts a statistical approach, where the probabilistic evolution is employed to iteratively reduce the impact of noise, ultimately revealing a cleaner underlying signal.

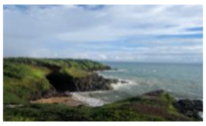
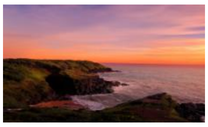
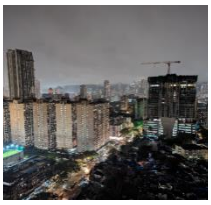
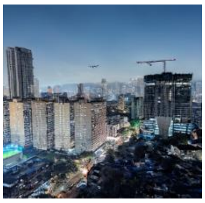
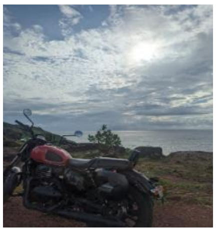
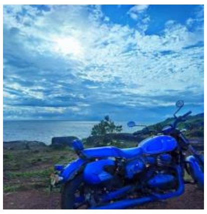
This approach is particularly valuable in scenarios where noisy data obscures the true patterns or information within a dataset. By leveraging the denoising diffusion probabilistic model, researchers and practitioners can systematically enhance the quality of the data, making it more amenable to subsequent analysis, interpretation, or utilization in applications such as image modification or other signal pro-cessing tasks.

6.7. Output Image

In this stage, the final output is generated, incorporating the modifications specified in the input text instruction. The image undergoes the prescribed changes as outlined in the textual guidance provided earlier in the process. This final output represents the result of the image modification project, reflecting the desired alterations, enhancements, or transformations dictated by the input text instruction. The successful generation of this output signifies the completion of the image modification process, where the integration of text instructions has influenced and shaped the visual characteristics of the modified image according to the project's objectives.

7. Results

Table 2. Outputs

Sr no.	Input Image	Text instruction	Output Image
1		Add a sunset	
2		Add a plane flying in the sky	
3		Make the bike blue	

In the given table we are to see the results obtained on a variety of photographs and images. We can successfully instruct the model by giving text input to commit changes to the original image. Our implementation is able to perform complex and challenging alterations to the image given by the user. It has the ability to add objects to the photographs and also change other properties such as its style and coloration. On row 1 of table 2 we are able to see the Input image of a mountain next to the sea and the text instruction “Add a sunset”, the output image changes the coloration of the sky to give it an orange horizon and ensures an appropriate amount of light is incident on the cliffside. On row 2 we are given the input image of an urban skyline with the text instruction “Add a plane flying in the sky”, the output image features a plane in the appropriate location but has the effect of slightly discoloring the sky. On row 3 the input image features 2 black swans on a lake and the edit instruction is “Make it into a watercolor painting”, the observed output changes the style of the photograph to make it look like it was a painting made with the help of watercolors. On row 4 we can see the input image includes a red bike with the sky as the backdrop and the edit instruction “Make the bike blue”, we can observe in the output image that the bike is turned blue but there is slight discoloration of the overall image. These results prove the capability and usability of the model in performing complex edits that users are able to do easily without prior knowledge in the field of editing with the use of our model.

7.1. Performance

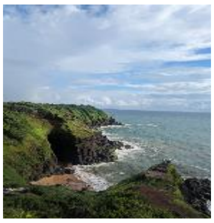
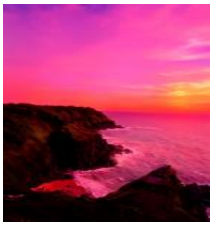
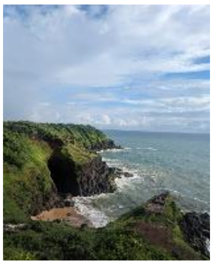
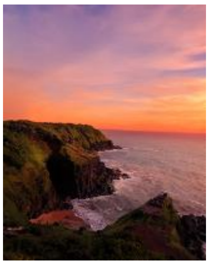
Table 3. Performance Table of Proposed Method

Number of iterations	Time	Memory (VRAM)
50	17 Minutes	10.4 GB
100	43 Minutes	10.7 GB

The above table offers insights into the computational performance of an innovative image modification model, showcasing the impact of varying iterations on time and memory usage. With 50 iterations, the image modification process took 17 minutes, utilizing 10.4 GB of VRAM. Increasing the iteration count to 100 extended the processing time to 43 minutes, accompanied by a slight VRAM rise to 10.7 GB. This highlights a discernible tradeoff: as users opt for more iterations, they can expect improved image quality, but at the expense of increased time and computational power. Therefore, the choice of iteration count becomes crucial, allowing users to balance the desire for enhanced image refinement with considerations for the associated time and resource costs.

7.2. Steps of Image

Table 3. Images generated at different Step

Input Image	Input Instruction	Number of Steps	Output Image
	add a sunset	50	
	add a sunset	100	

Here we can observe the overall upheaval of quality when the number of iterations is increased. The increase in the number of iterations has a negative effect on inference time of the model as we can see in table 2. The time taken for the inference is dependent on the hardware used to run the model. There is a point at which blindly increasing the number of iterations produces diminished return and takes much longer time without appropriate quality increase. The sweet spot for the number of iterations depends on the host hardware, input image and the complexity of the edit instruction. Hence, the output can vary in quality by running it on a preset number of iterations for different images.

As depicted in the provided table, the effect of increasing the number of iterations on the output image quality is apparent. However, it is imperative to note that this relationship is not linear; instead, it follows a pattern of diminishing marginal returns. Initially, with a relatively low number of iterations, significant improvements in image quality can be observed. This is because early iterations address prominent features or discrepancies in the input image, leading to noticeable enhancements.

Nevertheless, as the number of iterations continues to increase, the marginal improvements become less pronounced, eventually plateauing or even exhibiting a decline. This is attributed to several factors, including algorithmic convergence, computational overhead, and the inherent complexity of the image editing task. With each iteration, the algorithm converges towards an optimal solution, but the rate of convergence diminishes over time, leading to a slower rate of quality improvement.

8. Future Work

Amid rapid technological and global shifts, our future work is dedicated to innovation, equity, and sustainability, charting new directions for a better and more inclusive world.
Following is the planned future work of our project:
1. Adding factors to vary the output image
2. training with larger dataset
3. more steps to refine output
4. hosting on cloud
5. Developing an intuitive and user-friendly User Interface (UI) Establishing a connection between the UI and backend.

9. Conclusion

Image editing through text instruction represents a forefront in technological innovation, introducing a revolutionary approach to elevating digital images. Through the harmonious integration of natural language processing and advanced image processing techniques, this paradigm shift democratizes im-age manipulation, rendering it more accessible and intuitive for users with diverse skill levels. Beyond streamlining the editing process, this technology holds the potential for widespread application across various industries, fostering creativity and personalization in the digital landscape.

The effect on the final output by increasing the size of the dataset is substantial. The dataset used of size 436GB which contains 313010 image editing samples which allows for better image generation and cohesion. The effect of increase in number of steps is also noticeable but it suffers from diminishing returns. There is significant increase in time taken for inferencing as the number of steps increase. Due to the need for large amounts of video memory it is recommended to run the model on better hardware (>11 GB of video memory) which will give better inferencing time too.

Nevertheless, ongoing research endeavors and careful consideration of ethical implications are imperative for the responsible development and deployment of this technology. Ensuring inclusivity and addressing ethical concerns will be crucial to guarantee that image editing through text instruction benefits a broader audience in the evolving digital age. Striking a balance between innovation and responsible practices will be pivotal for the continued advancement and adoption of this transformative approach

References

[1] Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., ... & Irani, M.. Imagic: Text-based real image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)(pp. 6007-6017).Available from : https://doi.org/10.48550/arXiv.2210.09276

[2] Tsaban, L., & Passos, A. LEDITS: Real Image Editing with DDPM Inversion and Semantic Guidance. arXiv preprint arXiv:2307.00522 (2023). Available from : https://doi.org/10.48550/arXiv.2307.00522

[3] Brooks, T., Holynski, A., & Efros, A. A. Instructpix2pix: Learning to follow image editing instruc-tions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18392-18402) (2023).. Available from: https://doi.org/10.48550/arXiv.2211.09800

[4] Kawar, Bahjat, et al. "Imagic: Text-based real image editing with

diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. Available from : https://doi.org/10.48550/arXiv.2210.09276

[5] Miyake, Daiki, et al. "Negative-prompt Inversion: Fast Image Inversion for Editing with Text-guided Diffusion Models." arXiv preprint arXiv:2305.16807 (2023). Available from : https://doi.org/10.48550/arXiv.2305.16807

[6] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., … Simonyan, K. Flamingo: A visual language model for few-shot learning.(2022).Available from: https://doi.org/10.48550/arXiv.2204.14198

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34:8780–8794, 2021. Available from https://doi.org/10.48550/arXiv.2105.05233

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel CohenOr. An image is worth one word: Personalizing text-toimage generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022. Available from : https://doi.org/10.48550/arXiv.2208.01618

[9] Elharrouss, O., Almaadeed, N., Al-Maadeed, S., & Akbari, Y. Image inpainting: A review. Neural Processing Letters, 51, 2007-2028.(2020). Available from : https://doi.org/10.48550/arXiv.1909.06399

[10] Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., ... & Tang, J. Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems, 34, 19822-19835.(2021). Available from : https://doi.org/10.48550/arXiv.2105.13290

[11] Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., & Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors. In European Conference on Computer Vision (pp. 89-106). (2022, October). Cham: Springer Nature Switzerland. Available from : https://doi.org/10.48550/arXiv.2203.13131

[12] Wu, Q., Liu, Y., Zhao, H., Kale, A., Bui, T., Yu, T., ... & Chang, S. Uncovering the disentanglement capability in text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1900-1910). (2023). Available from : https://doi.org/10.48550/arXiv.2212.08698

[13] Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., & Shou, M. Z. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7452-7461). (2023) Available from : https://doi.org/10.48550/arXiv.2307.10816

[14] Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., ... & Guo, B. Vector quantized diffusion model for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10696-10706). (2022) Available from : https://doi.org/10.48550/arXiv.2111.14822

[15] Thao Nguyen, Yuheng Li, Utkarsh Ojha, Yong Jae Lee. Visual Instruction Inversion: Image Editing via Visual Prompting. ArXiv Preprint ArXiv:2307.14331. (2023, July 25) Available from : https://doi.org/10.48550/arXiv.2307.14331

[16] Li, J., Lu, W., Yang, M., Zhou, Y., & Yu, W. Text to Image Generation with Semantic-Spatial Aware GAN. ArXiv Preprint ArXiv:2104.00567. (2021, April 6) Available from : https://doi.org/10.48550/arXiv.2104.00567

[17] Gu, X., Yang, Y., Xu, H., Zhou, C., & Wang, Y. Text-Guided Neural Image Inpainting. ArXiv Pre-print ArXiv:2004.03212.(2020, April 7) Retrieved from Available from : https://doi.org/10.1145/3394171.3414017

[18] Dataset Link : https://instruct-pix2pix.eecs.berkeley.edu/clip-filtered-dataset/