AI-Powered Text-Guided Image Editing: Innovations in Fashion and Beyond

T. Charaa, T. Hamdeni, and I. Abdeljaoued-Tej

University of Carthage, ESSAI, Ariana, Tunisia
LR11ES13, FST, University Tunis-El-Manar, Tunisia
LR16IPT06, Institut Pasteur de Tunis, University Tunis-El-Manar, Tunisia

Abstract. Text-guided image editing on real images, particularly in the context of fashion, presents a highly versatile yet challenging task. This process requires that the editing system take as input only the original image and a textual instruction specifying the desired modifications. The system must autonomously identify the regions of the image to be altered while preserving the other characteristics of the original image. In this paper, we present our approach, which leverages state-of-the-art artificial intelligence techniques, including deep neural networks, large language models (LLMs), and advanced methods for image generation and editing, such as Stable Diffusion and InstructPix2Pix. By integrating these models, our system achieves precise interpretation of textual instructions and ensures consistent application of modifications while maintaining the visual integrity and authenticity of the original image. This framework provides a comprehensive and scalable approach for text-guided image editing, applicable to fashion and various other domains.

Keywords: Artificial Intelligence, Computer Vision, Image Editing, Neural Models, Text-Guided Image Editing, Deep Learning, Large Language Models (LLMs).

1 Introduction

Integrating text guidance into image processing tasks has become a major focus of computer vision research in today's digital age. The ability to interpret and execute image edits based on textual instructions represents a groundbreaking approach with wide-ranging applications, from creative content generation to image enhancement in various domains [4]. This work focused on developing a text-guided image editing system for fashion images, encompassing both realistic images sourced from catalogs or photographs and generated images. Unlike generic image editing, fashion-specific editing requires an exceptional level of detail to preserve the identity and characteristics of garments while performing precise modifications, such as changing colors, fabrics, or styles. Handling realistic images further adds to the complexity as current generative AI methods are predominantly optimized for synthetic image generation and often struggle with realistic inputs.
Our main goal is to create a system that can change images based on written instructions. We want to make it easier for people to edit images by using words instead of complicated editing tools.

Image editing has emerged as a powerful tool for users. The most advanced techniques available, especially in image editing nowadays are Text-to-Image Diffusion Models: Advanced tools such as DALL-E 2 [14] and Imagen [3] are capable of producing highly detailed images from textual descriptions. These models face challenges in fine-grained control, often struggling to capture complex scenarios, generate highly detailed images. However, those models can include difficulty capturing exact user intent or changing local adjustments to existing images, especially realistic ones.

David C. Wyld et al. (Eds): SOFTPA, NATAP, ARIN, CSITA, NMCO, CONEDU, ISPR – 2025
pp. 127-138, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.150611

In the other hand, ControlNets is an Attribute Editing with Diffusion Models [18, 19]. It allows for targeted modifications to existing fashion images; adjusting the color of a garment, adding embellishments, or changing the style of a shoe, all through text prompts. However, these edits are currently limited to pre-defined attributes and may struggle with preserving the original garment's identity during com- plex edits. The Object Addition and Editing are techniques like SDEdit and DiffEdit [5]. They allow for adding new fashion items or editing existing ones within an image. While these techniques offer exciting possibilities, they often rely on user-defined masks or noisy inversion processes, which can limit control over placement. Finally, Spatial Editing like DragGAN excel at efficiently moving existing fashion items within an image [11]. However, these techniques are currently limited to spatial manipulation tasks and cannot handle other editing needs like color changes or adding new items.

Despite these advancements, challenges remain. Many existing methods rely on optimization-based approaches, requiring per-prompt or per-image tuning, which hampers scalability and efficiency. Furthermore, a core requirement of image editing—preserving unedited regions while making semantic changes to relevant areas—remains inadequately addressed. Current methods often introduce unintended changes, compromising the quality and usability of the edited images.

A significant limitation of most existing generative methods is their poor adaptability to realistic images, as they are often optimized for synthetic data. Realistic fashion images present a unique challenge due to their complexity, requiring precise edits while preserving the texture, fit, and overall aesthetics of garments. Compounding this challenge, a lack of suitable datasets for training text-guided fashion image editing systems necessitated the development of custom data generation approaches.

To address the absence of a pre-existing dataset, we employed two novel approaches to generate the required training data. The first approach draws inspiration from the InstructPix2Pix [4] framework, where textual instructions were paired with corresponding edited images generated through iterative refinement. The second approach, based on our own research, leverages large language models (LLMs) to generate diverse and meaningful textual instructions. These instructions were then paired with images edited using Stable Diffusion models to simulate the desired outcomes. This dual strategy allowed us to construct a comprehensive dataset encompassing a wide range of editing scenarios, tailored specifically for the fashion domain.

In this work, we introduce a novel text-guided image editing system tailored specifically for fashion images. Our system bridges the gap between realistic and generated fashion images by offering precise and automated editing capabilities. Built upon the Stable Diffusion model, our approach ensures fidelity to the original image while enabling targeted edits guided by textual instructions. By addressing the limitations of previous methods and overcoming the dataset challenge, we demonstrate a significant advancement in handling realistic images, which has been a critical bottleneck in the field. This document is structured to first present an overview of the business context and dataset preparation, focusing on integrating diverse inputs of both realistic and generated fashion images. We then detail the development of our system, including its architecture and optimization strategies, and demonstrate significant improvements over existing methods in the fashion domain.

2 Background

The key feature of transformer-based architectures is the attention mechanism, which captures intricate dependencies within input and output sequences. This innovation has been pivotal in adapting these models for multimodal tasks such as text-guided image editing, enabling them to revolutionize not only natural language processing (NLP) [1,15] but also image generation and editing by effectively bridging textual and visual modalities. Although traditional attention mechanisms were combined with recurrent networks to improve sequence modeling [10]. However, this approach poses challenges, particularly regarding computational efficiency and handling long dependencies. The transformer model emerges as a state-of-the-art solution in the realm of NLP, specifically designed to address these challenges. This innovative approach revolutionizes the field of NLP, offering unparalleled efficiency and effectiveness in capturing intricate relationships within sequences. It aims to revolutionize image editing by offering a user-friendly alternative to complex editing tools. By enabling users to articulate desired modifications through instructions, we aim to streamline the editing process significantly.

Over recent years, generative models have advanced to produce high-quality synthetic images by employing techniques such as Denoising Diffusion Probabilistic Models (DDPM) [6]. These models simulate a diffusion process where noise is progressively added to an image and then systematically removed, allowing for the reconstruction of clear, realistic images and the generation of new visuals that reflect patterns from the training data.

Diffusion models have become pivotal in image editing applications [16, 17], excelling in tasks such as denoising, inpainting, super-resolution, and style transfer [2,8]. Their adaptability makes them particularly suitable for integrating textual conditions as guidance, enabling intuitive user interaction through simple text inputs. By eliminating the need for expertise in traditional editing tools, text-based guidance opens up accessibility to a wider audience. Although these methods show promise in image editing, they often struggle with synthesizing image information and targeted editing tasks due to gaps between image modalities and the limited context provided by textual inputs. Text-based instructions usually do not convey the full context of the image, which makes it difficult to accurately determine the key elements that need to be changed. Consequently, the edits might not align with the user's expectations.

To overcome these challenges, additional guidance techniques have been explored. One common approach is to incorporate manual annotations, such as masks, to explicitly mark the regions of an image to be edited [12]. This method enhances precision by directing the model to focus on specific areas based on user-customized instructions. While effective, the reliance on manual annotations limits scalability and reduces the general applicability of text-guided systems for large-scale or automated editing tasks.

The emergence of multimodel learning models such as CLIP Contrastive Language-Image Pre-training [13] represents a significant advancement in the field. CLIP, a neural network developed by OpenAI, is designed to learn visual concepts from language descriptions. It addresses the challenge of learning visual representation from natural language by adapting a contrastive learning approach. Its strength lies in its ability to leverage a diverse and expansive dataset of both images and text. Contrastive learning, a widely used technique in machine learning—particularly in unsupervised learning—focuses on training AI models to identify similarities and differences across various data points.

Our objective focuses on building a system that can effectively edit images based on user-provided textual instruction. We used a pre-trained model that can understand these instructions and use them to edit the image. To enhance the system's performance for our specific needs, we finetuned the model using a collection of text instructions along with their corresponding original and edited images. This helped the system work faster and more accurately, while also making sure it understands the instructions we give it.

3 Data and Methods

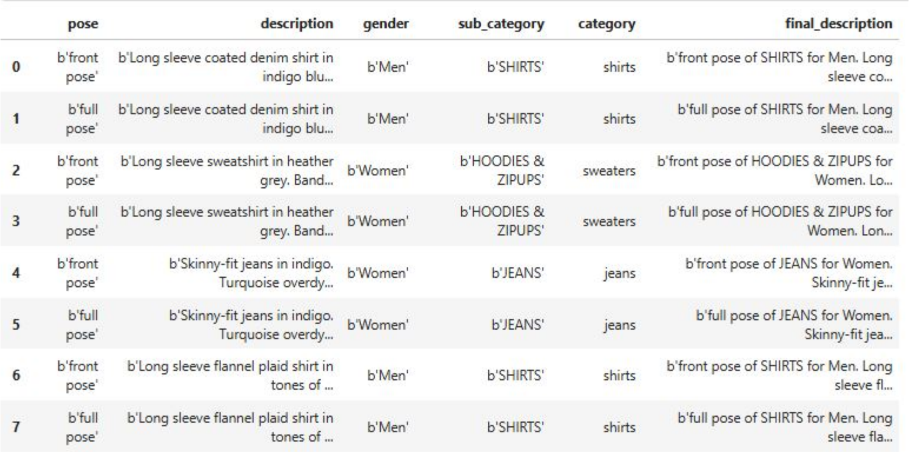
Developing an innovative image editing system driven by textual instructions requires not only advanced models but also a robust and representative dataset. Such a dataset must effectively bridge the gap between natural language directives and the desired visual modifications, especially in the domain of fashion imagery. However, existing datasets tailored for this specific task are scarce, posing a significant challenge. To overcome this, we devised two complementary approaches for generating a dataset specifically tailored to our use case. These approaches were informed by extensive research in the field and were designed to balance realism, diversity, and relevance to fashion image editing.

3.1 Approach 1: Following the InstructPix2Pix Framework

This approach is inspired by the InstructPix2Pix framework, which combines pre-trained language and image generation models to create paired datasets. This method relies on the Fashion-Gen dataset, a large collection of fashion images with textual descriptions.

FashionGen dataset: The Fashion-Gen dataset is a large-scale dataset containing high-resolution fashion images (1360×1360 pixels) with detailed textual descriptions written by professional stylists. This dataset is widely used in fashion image generation, retrieval, and text-guided modifications.

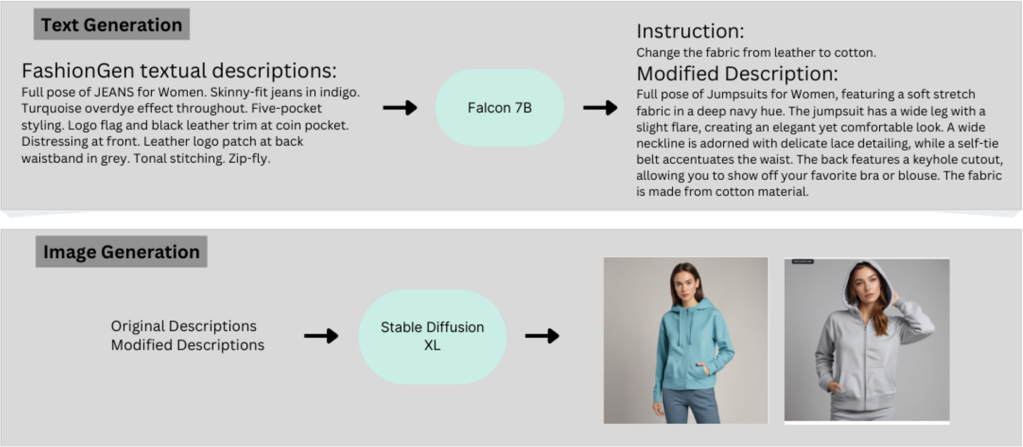
The dataset consists of 293,008 fashion images, each captured from multiple angles (e.g., front pose, full pose). These images are paired with textual attributes describing the clothing's category, style, material, and color. Below is an image sample from the Fashion-Gen dataset, illustrating the pose variations, descriptions, and final preprocessed descriptions:



	pose	description	gender	sub_category	category	final_description
0	b'front pose'	b'Long sleeve coated denim shirt in indigo blu...	b'Men'	b'SHIRTS'	shirts	b'front pose of SHIRTS for Men. Long sleeve co...
1	b'full pose'	b'Long sleeve coated denim shirt in indigo blu...	b'Men'	b'SHIRTS'	shirts	b'full pose of SHIRTS for Men. Long sleeve coa...
2	b'front pose'	b'Long sleeve sweatshirt in heather grey. Band...	b'Women'	b'HOODIES & ZIPUPS'	sweaters	b'front pose of HOODIES & ZIPUPS for Women. Lo...
3	b'full pose'	b'Long sleeve sweatshirt in heather grey. Band...	b'Women'	b'HOODIES & ZIPUPS'	sweaters	b'full pose of HOODIES & ZIPUPS for Women. Lon...
4	b'front pose'	b'Skinny-fit jeans in indigo. Turquoise overdy...	b'Women'	b'JEANS'	jeans	b'front pose of JEANS for Women. Skinny-fit je...
5	b'full pose'	b'Skinny-fit jeans in indigo. Turquoise overdy...	b'Women'	b'JEANS'	jeans	b'full pose of JEANS for Women. Skinny-fit jea...
6	b'front pose'	b'Long sleeve flannel plaid shirt in tones of ...	b'Men'	b'SHIRTS'	shirts	b'front pose of SHIRTS for Men. Long sleeve fl...
7	b'full pose'	b'Long sleeve flannel plaid shirt in tones of ...	b'Men'	b'SHIRTS'	shirts	b'full pose of SHIRTS for Men. Long sleeve fla...

Fig. 1. Sample Data from the Fashion-Gen Dataset

To generate modification instructions, we used the Falcon 7B language model, while a pre-trained Stable Diffusion model was applied to create corresponding edited images. The dataset was then constructed by pairing the original images with their modified versions based on text descriptions.
The overall workflow for this approach is summarized in Figure 2, which outlines the step-by-step process to generate high-quality paired data for training.



Text Generation

FashionGen textual descriptions:
Full pose of JEANS for Women. Skinny-fit jeans in indigo. Turquoise overdye effect throughout. Five-pocket styling. Logo flag and black leather trim at coin pocket. Distressing at front. Leather logo patch at back waistband in grey. Tonal stitching. Zip-fly.

Instruction:
Change the fabric from leather to cotton.

Modified Description:
Full pose of Jumpsuits for Women, featuring a soft stretch fabric in a deep navy hue. The jumpsuit has a wide leg with a slight flare, creating an elegant yet comfortable look. A wide neckline is adorned with delicate lace detailing, while a self-tie belt accentuates the waist. The back features a keyhole cutout, allowing you to show off your favorite bra or blouse. The fabric is made from cotton material.

Image Generation

Original Descriptions
Modified Descriptions

Fig. 2. Diagram of the first approach

Limitations of Approach 1

While this approach offered a scalable method for generating a large dataset, it presented certain limitations in our case:

- Ambiguity in Instruction Generation: Falcon 7B often produces general or vague instructions, making precise modifications difficult.
- Inconsistency in Image Edits: The Stable Diffusion model lacks fine control over localized edits, sometimes introducing unintended changes that distort the original image.
- Challenges with Realistic Edits: Due to the diverse nature of the Fashion-Gen dataset, maintaining consistency between edited and original images is difficult, especially when handling fine-grained modifications such as fabric texture or subtle color variations.

These limitations reduce the reliability of the dataset for training high-quality models. To overcome these issues, we developed a second approach that ensures better precision and control over the editing process.

3.2 Approach 2: Controlled Edits Using an Inpainting Model

Here, we explored and proposed an approach for generating the dataset. The goal of this new method is to ensure that the generated images for the original and modified descriptions maintain a more faithful visual relationship. To achieve this, we utilized a pre-trained Stable Diffusion inpainting model [9], which allows for more controlled and precise edits to specific regions of an image based on given masks. This approach leverages Human-parsing-dataset and Fashion-controlnet-dataset from HuggingFace [7], which includes images, masks, and Captions: Human-parsing-dataset typically contains labeled images in which each pixel is annotated to correspond to a specific category, such as body parts (e.g., head, arms, legs), clothing (e.g., shirts, pants), or accessories. This dataset is used in tasks like person segmentation and human part recognition to enable more detailed image understanding. The Fashion ControlNet Dataset is designed for fashion image generation and manipulation tasks. It includes fashion images annotated with segmentation maps, and other attributes that allow models to generate or edit fashion images based on input prompts. The dataset facilitates advanced applications in fashion image synthesis. By using the inpainting model, we can focus on making localized edits that are directly related to the modified descriptions, ensuring that the rest of the image remains visually consistent with the original. In this section, we provide a detailed explanation of the code of the initiative approach used for data generation.

Masks and Impainting Model

The mask is a crucial input for the inpainting model, guiding which parts of an image should be edited. Typically represented as a binary image, the mask highlights areas for modification in white while keeping unchanged regions black. The model utilizes these masks, derived from datasets like the human-parsing and Fashion-controlnet datasets, to accurately target specific garments or body parts for editing. By employing a pre-trained stable diffusion inpainting model, the system can perform localized edits without generating a new image from scratch. This approach ensures that modifications integrate seamlessly with the original image, preserving visual consistency by maintaining the structure and appearance of unchanged areas. Overall, this method allows for precise adjustments while keeping the integrity of the image intact.

Labeling the Dataset

Detect garment from the garments list and apply into a new column in modified-dataset called detected-garment, where each caption in clip-captions is processed to detect a garment.

```python
modified_dataset = pandas.DataFrame({'CLIP_captions': dataset['CLIP_captions']})

def detect_garments_in_caption(caption, garments):
    for garment in garments:
        if garment in caption.lower():
            return garment
    return None

modified_dataset['detected_garment'] = modified_dataset['CLIP_captions'].apply(lambda caption: detect_garments_in_caption(caption, garments)
)
```

Listing 1.1. Detecting the clothing item

Identifies and handles rows with missing garment detection, it is then removed from the modified-dataset but resets the original image-index after dropping them.

Instruction Generation

Color-names is a list of a predefined color. Get-instruction-format is a function that applies a random instruction for changing the garment's color by choosing one of three predefined formats. Pil-to-bytes is a function to convert a PIL image object into byte data, which is used for saving the image in the specific format (PNG).[1]

[1] We applied the same logic for the fabric changes, where the function selects from a predefined fabric list and constructs instructions.

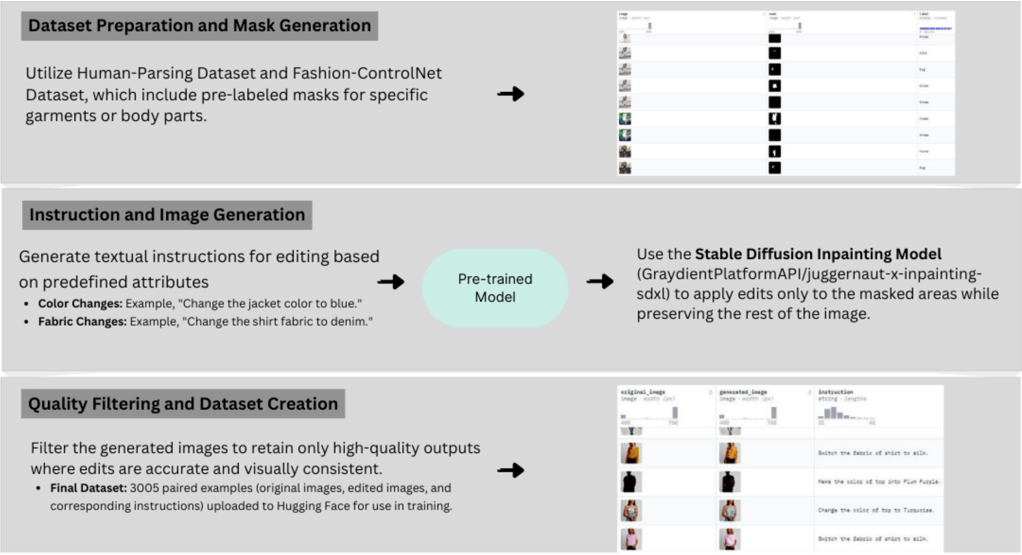
```python
color_names = ['Red', 'Orange', 'Yellow', 'Green', 'Blue', 'Purple', 'Pink'
    ]
def get_instruction_format(label, selected_color):
    formats = [
        f"Change the color of {label} to {selected_color}.",
        f"Replace the {label} color with {selected_color}.",
        f"Make the color of {label} into {selected_color}.",
    ]
    return random.choice(formats)

def pil_to_bytes(img):
    img_byte_arr = io.BytesIO()
    img.save(img_byte_arr, format='PNG')
    img_byte_arr = img_byte_arr.getvalue()
    return img_byte_arr
```

Listing 1.2. Detecting the clothing item

We apply the same logic for the fabric changes, where the function selects from a predefined fabric list and constructs instructions.

Image Generation

For the image generation process, we used the inference of the pre-trained inpainting model `GraydientPlatformAPI/jux-inpainting-sdxl` to produce the final dataset. After apply this model on the human-parsing-dataset and Fashion-controlnet-dataset, we filtered the results to get a set of 3005 high-quality images. The resulting dataset contains original images, edited images, and the desired instructions. Fig.3 illustrates the complete pipeline of this approach.



Dataset Preparation and Mask Generation

Utilize Human-Parsing Dataset and Fashion-ControlNet Dataset, which include pre-labeled masks for specific garments or body parts.

Instruction and Image Generation

Generate textual instructions for editing based on predefined attributes

- **Color Changes:** Example, "Change the jacket color to blue."
- **Fabric Changes:** Example, "Change the shirt fabric to denim."

Use the **Stable Diffusion Inpainting Model** (GraydientPlatformAPI/juggernaut-x-inpainting-sdxl) to apply edits only to the masked areas while preserving the rest of the image.

Quality Filtering and Dataset Creation

Filter the generated images to retain only high-quality outputs where edits are accurate and visually consistent.

- **Final Dataset:** 3005 paired examples (original images, edited images, and corresponding instructions) uploaded to Hugging Face for use in training.

Pre-trained Model
original_image
generated_image
instruction
Switch the fabric of shirt to silk.
Make the color of top into Plum Purple.
Change the color of top to Turquoise.
Switch the fabric of shirt to silk.

Fig. 3. Diagram of the second approach

3.3 Fine-tuning Experiments

We explored the fine-tuning experiments conducted on InstructPix2Pix using our generated dataset to enhance its in text-guided fashion image editing [4]. Utilizing the Freezing method for fine-tuning, we adapted the model to apply precise modifications. By freezing most of the model's layers and training only a subset of parameters, we ensured efficient fine-tuning with minimal resource usage.

During fine-tuning the InstructPix2Pix model, several arguments are used to control the training process: The model processes 16 samples before updates; the number of samples used during the validation process at each step is equal to 4; the models iterate the entire dataset 50 times during training; the use of the 8-bit Adam optimizer, which reduces memory usage by quantizing Adam's weights to 8-bit precison.

Additionally, to monitor the fine-tuning process, we integrated Weights and Biases (WandB) to track key training metrics in real time. This platform detailed insights into the model's performance, including loss curves, parameter updates, and relevant metrics, ensuring we could adjust the training process for optimal results.

```
vae.requires_grad_(False)
text_encoder.requires_grad_(False)
```

Listing 1.3. Freezing Code Block

There are two distinct approaches to fine-tuning the InstructPix2Pix. Both strategies aim to improve the models' ability to generate concise images, but they differ in their methodology. Additionally, we will apply these approaches using 95% of the data for the training and 5% for the validation. The freezing methodology is a widely used technique in finetuning, which involves locking specific layers of the model to prevent their weights from being updated during training. This approach enhances the efficiency of the finetuning process and helps reduce the risk of overfitting.

To better illustrate the dataset distribution and fine-tuning results, we present Table 1, which outlines the number of samples used for training and validation, along with the average SSIM scores achieved during training and testing.

Table 1. Dataset Distribution and SSIM Scores for Fine-Tuning

Dataset Split	**Total Samples**	**Percentage**	**SSIM (Training Avg.)**	**SSIM (Testing Avg.)**
Training Set	2855	95%	0.87	-
Validation Set	150	5%	-	0.89

Bitsandbytes is a popular approach and library for optimizing model fine-tuning, particularly when using large models. It focuses on reducing the memory footprint of models by utilizing 8-bit and 4-bit precision rather than the usual 16-bit or 32-bit of precision. This allows users to fine-tune large models even on hardware with limited memory, such as GPUs or multi-GPU setups.

3.4 Evaluation

In this part, we evaluate the performance of our model using the Structured Similarity Index Measure (SSIM), which is a metric used to evaluate the quality of images by comparing the structural information between the original edited image (the edited image in the dataset) and the generated image during the training process. It is designed to assess image quality based on three key factors: Luminance measures the brightnes of the images; Contrast measures the contrast in intensity; Structure measure the spatial arrangement of pixels. The SSIM index produces values ranging from -1 to 1, where 1 indicates perfect structural similarity between the two compared images.

The SSIM index is defined in Eq. (1):

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{1}$$

where μ_x and μ_y are the mean values of the two images x and y; σ_x^2 and σ_y^2 are the variances of the images x and y; σ_{xy} refers to the covariance between the two images; $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$ are constants to stabilize the division with small denominators, where L is the dynamic range of the pixel values (typically 255 for 8-bit images), and K_1 and K_2 are small constants (often set to 0.01 and 0.03, respectively).

4 Results and Discussion

Starting with a lower value around 0.75, the model quickly learns and improves over the first few thousand steps (see Fig. 4). By the end of the training, the SSIM approaches 0.9, indicating that the model has achieved high structural similarity between its outputs and the target images. The validation SSIM curve exhibits a similar trend to the training SSIM curve, serving as a strong indication that the model generalizes effectively to unseen data. Starting near 0.76, the score increases rapidly in the early training phases and continues to improve steadily. By the end of training, the validation SSIM also approaches 0.9, showing that the model maintains high performance on the validation set, avoiding overfitting. This parallel movement between training and validation SSIM curves suggests balanced learning and a robust model.

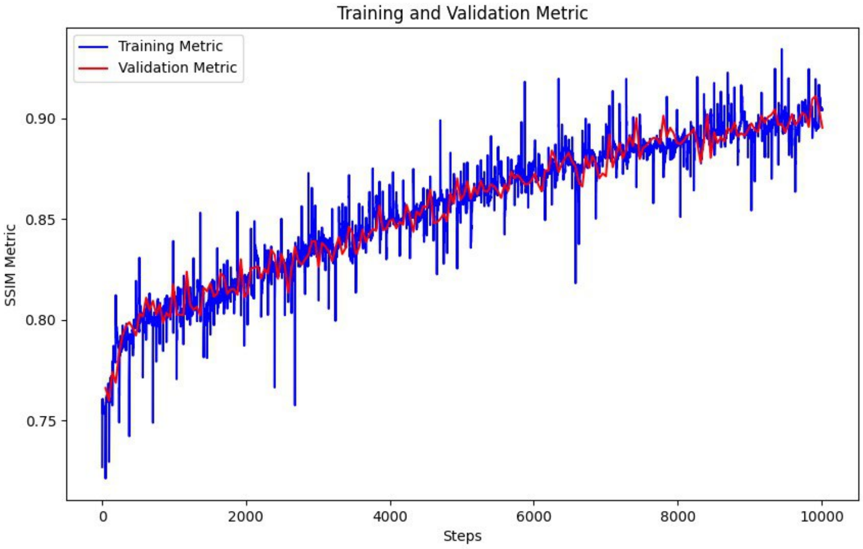


Fig. 4. Training and validation SSIM curve

Our study presents a significant advancement in text-guided image editing within the fashion domain. The high Structural Similarity Index Measure (SSIM) scores, approaching 0.9, indicate that our model effectively maintains the structural integrity of the original images while applying precise edits based on textual instructions. This demonstrates the model's capability to understand and implement user intentions, which is crucial for applications requiring detailed and accurate image modifications.

Compared to existing models like InstructPix2Pix and Stable Diffusion, our approach addresses key challenges identified in prior research. InstructPix2Pix [4], while innovative, struggled with maintaining image consistency and specificity in instruction interpretation, especially with realistic images. Similarly, Stable Diffusion models often lacked precise control over localized edits, leading to unintended changes. Our method, leveraging controlled inpainting techniques and fine-tuning with a custom dataset, provides enhanced precision and consistency. This aligns with recent advancements in diffusion models that emphasize conditional control for improved editing fidelity [18].

5 Conclusion

We have made significant strides in the field of text-guided image editing, particularly in modifying the color and fabric of garments while preserving the integrity of the original images. Our system excels at delivering precise results for individual element modifications, though it currently faces limitations in adding or removing objects. One notable challenge is in fabric editing, where changes sometimes inadvertently affect the garment's color. Addressing this issue will be a primary focus in our future developments to improve the system's accuracy and reliability.

Moreover, expanding the model's capabilities to accommodate more complex edits will unlock new possibilities for advanced image manipulation, allowing users greater creative freedom. This work not only highlights the potential of contemporary models in image editing but also establishes a foundation for future innovations in this domain. The methodologies and techniques developed here can be further refined and adapted, paving the way for automated, user-driven image transformations on a larger scale, ultimately enhancing the user experience and broadening the scope of digital creativity.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* (Vol. 30). DOI: 10.5555/3295222.3295349.
2. Avrahami, O., Lischinski, D., and Fried, O. (2022). Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18208-18218). DOI: 10.1109/CVPR52688.2022.01767.
3. Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Chan, K., et al. (2024). Imagen 3. *CoRR.* arXiv:2408.07009.
4. Brooks, T., Holynski, A., and Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18392-18402). DOI: 10.1109/CVPR52688.2023.01839.
5. Couairon, G., Verbeek, J., Schwenk, H., and Cord, M. (2023, May). DiffEdit: Diffusion-based Semantic Image Editing with Mask Guidance. In 11^{th} *International Conference on Learning Representations.* DOI: 10.1109/ICLR.2023.00001.
6. Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, (pp. 6840-6851). DOI: 10.5555/3454287.3454870.
7. Jain, S. M. (2022). Hugging face. In Introduction to transformers for NLP: With the hugging face library and models to solve problems (pp. 51-67). Springer. DOI: 10.1007/978-1-4842-8844-3_2.
8. Kawar, B., Zada, S., et al. (2023). Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6007-6017). DOI: 10.1109/CVPR52688.2023.00607.
9. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11461-11471). DOI: 10.1109/CVPR52688.2022.01117.
10. Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452, 48-62. DOI: 10.1016/j.neucom.2021.03.091.
11. Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., and Theobalt, C. (2023, July). Drag your gan: Interactive point-based manipulation on the generative image manifold. *In ACM SIGGRAPH 2023 Conference Proceedings* (pp. 1-11). DOI: 10.1145/3588432.3591500.
12. Patashnik, O., Garibi, D., Azuri, I., Averbuch-Elor, H., and Cohen-Or, D. (2023). Localizing object-level shape variations with text-to-image diffusion models. In P*roceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 23051-23061). DOI: 10.1109/ICCV51070.2023.02107.
13. Radford, A., Kim, J. W., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR. DOI: 10.48550/arXiv.2103.00020.
14. Reddy, M. D. M., Basha, M. S. M., Hari, M. M. C., and Penchalaiah, M. N. (2021). Dall-e: Creating images from text. *UGC Care Group I Journal*, 8(14), 71-75. DOI: 10.5281/zenodo.5790549.
15. Rivas, P., and Zhao, L. (2024). Marketing and AI-Based Image Generation: A Responsible AI Perspective. In *International Conference on Information and Communication Technology for Intelligent Systems* (pp. 141-151). Singapore: Springer Nature Singapore. DOI: 10.1007/978-981-97-5810-4_13.
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695). DOI: 10.1109/CVPR52688.2022.01042.
17. Song, J., Meng, C., and Ermon, S. (2021). Denoising Diffusion Implicit Models. In *International Conference on Learning Representations.* DOI: 10.48550/arXiv.2010.02502.
18. Zhang, L., Rao, A., and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3836-3847). DOI: 10.1109/ICCV56361.2023.00384.
19. Zavadski, D., Feiden, J. F., and Rother, C. (2025). ControlNet-XS: Rethinking the Control of Text-to-Image Diffusion Models as Feedback-Control Systems. In *European Conference on Computer Vision* (pp. 343-362). Springer, Cham.

Authors

Tasnim Charaa graduated from the Engineering School of Statistics and Information Analysis at the University of Carthage, Tunisia. She specializes in Generative Artificial Intelligence and Data Analysis, focusing on developing innovative solutions that leverage data for enhanced decision-making. With a keen interest in the applications of artificial intelligence, she explores how generative models can be utilized to create predictive analytics and improve data-driven strategies.
tasnim.charaa@essai.ucar.tn

Tasnime Hamdeni received a PhD in Software Engineering and Mathematics from the University of Toulon in France and a PhD in Applied Mathematics from ENIST, University of Tunis El-Manar in Tunisia. She completed her Engineering degree in Applied Mathematics and Modeling at ENSIT, University of Tunis. Currently, she is an Assistant Professor at the Engineering School of Statistics and Information Analysis at the University of Carthage, Tunisia. Her research interests include Signal, Image and Artificial Intelligence.
tasnim.hamdeni@essai.ucar.tn; https://orcid.org/0000-0002-8742-3641

Ines Abdeljaoued-Tej received her PhD from Sorbonne University, following a Bachelor's degree in Pure Mathematics and a Master's degree in Algorithmics. Currently, she is an Assistant Professor at the Engineering School of Statistics and Information Analysis at the University of Carthage in Tunisia. Her research interests include Bioinformatics, Artificial Intelligence, and Symbolic Computation.
ines.tej@essai.ucar.tn; https://orcid.org/0000-0002-1796-7897