Fashion Is All You Need: Generating Clothing Styles Using Diffusion

Adeesh Bhargava
Electrical and Computer Engineering
adeeshb@andrew.cmu.edu

Anvesh Reddy Gummi
Mechanical Engineering
agummi@andrew.cmu.edu

Kriti Kukreja
Information Networking Institute
kkukreja@andrew.cmu.edu

Meghana A Rajeev
Artificial Intelligence and Innovation
mrajeev@andrew.cmu.edu

1 Abstract

The fashion industry stands to benefit greatly from the use of AI-generated fashion trends, which can provide personalized fashion recommendations based on an individual's body type, skin tone, and personal preferences, ultimately transforming the way people perceive and approach fashion. Our work focuses on generating pose-aware, photo-realistic clothes for a target image, which takes into account the visual context to ensure accuracy and aesthetic appeal. We use a combination of cloth segmentation, augmentation, and a stable diffusion model that considers the context of the prompt to generate high-quality results. We also utilize PIDM (Person Image Synthesis via Denoising Diffusion Model)(1) to generate images of the person wearing the clothing item in different poses to ensure proper alignment with the body in the target image. Our approach produces realistic-looking clothing items that accurately reflect the texture of the input image while maintaining the integrity of the target image's pose and bodily features.

We provide our working code here.

2 Motivation

The use of AI-generated fashion trends is beneficial for the fashion industry, and can also revolutionize the way individuals perceive and approach fashion. With AI technology, individuals can receive personalized fashion recommendations based on their body type, skin tone, and personal preferences. This can save individuals time and effort in finding the right clothes that fit their style and suit their body type better. This can help break down traditional fashion norms and inspire people to express their individuality and creativity through their clothing choices.

Overall, the integration of AI technology in the fashion industry has the potential to benefit both designers and individuals alike. With the ability to generate new and diverse fashion trends, AI can help reduce waste, promote inclusivity and diversity, and empower individuals to experiment with their style, ultimately transforming the way we perceive and approach fashion.

3 Experiments

Our work for the midterm report has been focused on researching and fine tuning the various controlled diffusion models that exist. We have attempted to extensively visualise the capabilities and shortcomings of the current SOTA models listed below, specifically for our downstream task, which is to generate photo-realistic clothes for a given target test subject:

- RePaint(2): The RePaint model is great at using diffusion to in-paint the missing pixels in an image. Our preliminary approach was to mask out the clothes of a person in a target image using segmentation for clothes and then run the enhanced DDPM diffusion model (3) using class control of human faces. The model is finetuned to generating facial features but still lacks context and control over the generated results.

(a) *Input Image* (b) *Segmentation Mask* (c) *Input + Mask* (d) *Reconstruction*

Figure 1: *RePaint: Reconstructions using 256x256 Classifier Dataset Model*

We observed the following shortcomings for our specific use-case:

1. The model is heavily dependent on the class control (Faces, CelebHQ-A, "Clothes" when fine tuned with custom dataset) and the segmentation mask provided.
2. The clothes generated are a linear interpolation and biased on the custom dataset and need some more prompting for the results to look more desirable and meaningful.

- Stable Diffusion (4): The stable diffusion v1.4 model weights are openly available which we tried to use for our downstream task. The Vanilla Stable diffusion model takes as inputs: an input image and a text prompt and generates as output the intended image. Here are a few of our experiments (best results):

(a) *Prompt: cyberpunk techwear streetwear look and clothes, we can see them from feet to head, highly detailed and intricate, golden ratio, beautiful bright colors, hypermaximalist, futuristic, cyberpunk setting, luxury, elite, cinematic, techwear fashion, Errolson Hugh, Sacai, Nike ACG, Yohji Yamamoto, Y3, ACRNYM, matte painting –w 2176 –h 3840 –iw 1*

(b) *Prompt: beautifully lit fashion portrait of black female marble statue with symmetrical face, the statue is wearing huge oversize quilted flowing floor length long puffer jacket by balenciaga, yeezy, y 3, yohji yamamoto, comme de garcon, rei kawakubo, drape, sharp focus, clear, detailed,, romantic, brutalist concrete architecture in the background,fashion, magazine shoot, glossy*

Figure 2: *Potential of Stable diffusion with elaborate text prompts to generate high quality results for clothing and fashion.*

Therefore , from the above results , we can infer that the stable diffusion model and dataset is a great start point for our use-case and has good potential and input dataset to check the feasibility of our idea and pipeline. The stable diffusion v1.4 model is used for all further models and fine-tunings or customizations.

(a) *Prompt: Change the shirt of the man to a colourful tee or T-shirt*

(b) *Prompt: Change the shirt of the man to a sweatshirt without changing his face*

Figure 3: *Vanilla Stable Diffusion- Image + Text Prompt Only*

We observed the following shortcomings for our specific use-case:

1. The Vanilla Stable diffusion even if fine-tuned does not produce controlled results without a mask and significant amounts of prompt engineering to get desired results.
2. Facial Features are often distorted in the resulting image without the mask.
3. Even if good results are generated, the body shape and posture information is generalised to source images in the training dataset, generating photo-realistic but visibily fake and morphed images.
4. The context in the prompt is often not sufficient to direct the model into generating the intended images.
5. Examples: Applying colour to specific regions only, features that cannot be explained in text embeddings like texture(dotted/printed), material (silky,coarse) and background(summer shirt, winter jacket etc) are lacking in current approach and architectures.

- Stable Diffusion + Segmentation Mask + Prompt Engineering: To test the maximum capabilities of the stable diffusion model for our task, we tried to fine tune the model pipeline adding a segmentation mask and providing most discrete, elaborate and effective prompts to generate following results.

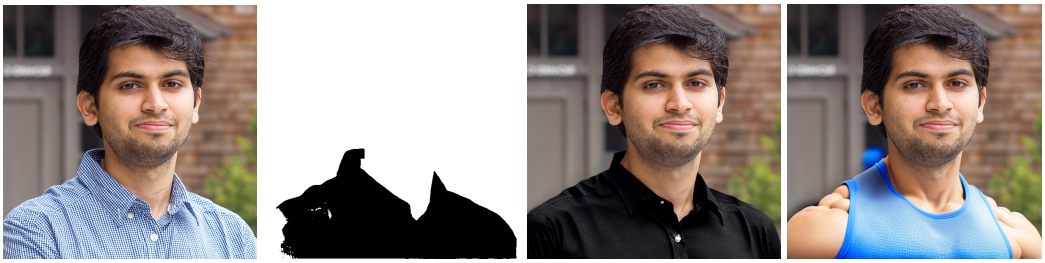
(a) *Prompt: Change the clothes of the woman in the photo to a brown western woman's dress*

(b) *Prompt: Change the clothes of the man to a light blue casual sweatshirt*

(c) *Prompt: Change the clothes of the woman to wear a light blue casual formal jacket on a t-shirt*

(d) *Prompt: Change the clothes of the man to wear black formals with a tie and a black men's suit*

Figure 4: *Stable Diffusion Results with Image + Mask + Text Prompt Engineering Pipeline and Fine tuning hyper parameters*



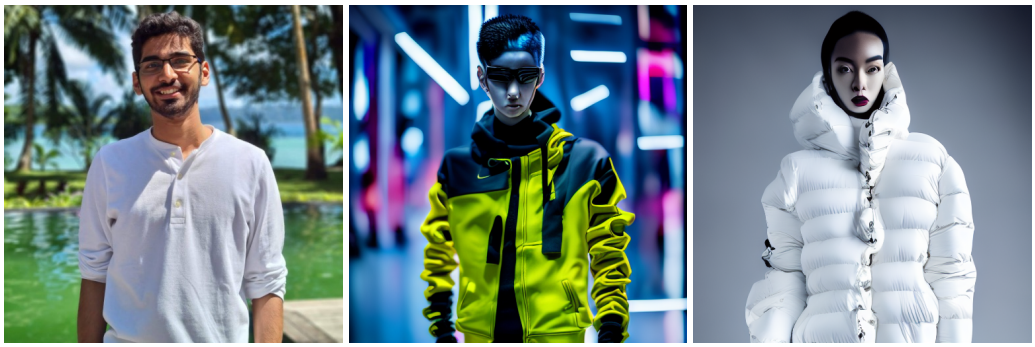
(a) *Input image and segmentation mask used for the generation process*

(b) *Prompt: Change the shirt of the man to various outfits*

Figure 5: *Stable Diffusion- Image + Text + Mask Shortcomings*

The model seems to create good results with effective prompt engineering, however this approach still has the following shortcomings:

1. The pose and body structure of the input image is not preserved.
2. The outputs generated are using the body shapes of the input dataset.
3. The model cannot incorporate knowledge embedded in the text without more supervision or context, resulting in elaborate prompts and having to do a lot of trial and error to make the results seem realistic.
4. Model cannot generate texture ,text, positioning, background affecting clothing style etc which is not understood by textual context alone.



(a) (b) (c)

Figure 6: *(a) Reference Image (b) Prompt:cyberpunk techwear streetwear look and clothes, we can see them from feet to head, highly detailed and intricate, golden ratio, beautiful bright colors, hypermaximalist, futuristic, cyberpunk setting, luxury, elite, cinematic, techwear fashion, Errolson Hugh, Sacai, Nike ACG, Yohji Yamamoto, Y3, ACRNYM, matte painting (c) Prompt:lit fashion portrait of him wearing huge oversize quilted flowing floor length long puffer jacket by balenciaga, yeezy, y 3, yohji yamamoto, comme de garcon, rei kawakubo, drape, sharp focus, clear, detailed, detailed, white, symmetrical, vogue, editorial, fashion, magazine shoot, glossy*

- InstructPix2Pix (5): InstructPix2Pix is an image-to-image model that takes in a reference image and text as input and outputs an edited image according to the text instruction. We used the pre-trained model off the shelf and experimented with it. We provided a reference image and text instructions as shown in the fig.

  We noticed that though it generated realistic images according to the prompt, the reference human model we provided as input completely changed. Another human model was generated wearing the instructed clothes.

- DreamBooth (6): DreamBooth is an image-to-image model that generates personalized fashion images based on instance prompts and class prompts. To fine-tune the model for

our specific use case, we provided four instance images of a person to give the model information about their body shape, size, and preferred fit. Additionally, we selected nine class images featuring different celebrities wearing various types of clothing to provide the model with examples of desired clothing styles, colors, and materials. The instance prompt we gave was "photo of Adeesh human" and the class prompt we gave was "a photo of human". However, the photos generated by DreamBooth were mostly distorted, and the clothes in the images did not even resemble actual clothes.

(a) *Prompt: A photo of Adeesh wearing a red sweater* (b) *Prompt: A photo of Adeesh wearing a pink shirt.*

Figure 7: *Images generated by Fine Tuned DreamBooth.*

- DiffFashion: It is a fine-tuned version of the Stable Diffusion model OpenJourney. OpenJourney in turn is a Stable Diffusion model trained on Midjourney images. For Diffusion_Fashion, the model is fine tuned on Fashion Product Images Dataset, a 25GB Kaggle dataset. It can only take text prompts. The following are some observations about the model(Fig. 7):

    1. The model correctly generates clothes in the material that is asked.
    2. When no color is mentioned, the model sometimes generates greyscale images.
    3. Faces are usually distorted, except when the model is specifically prompted to generate symmetrical faces.
    4. The current limitations of the model require a large number of prompts to produce outputs that align with our desired outcome. However, this approach is not optimal for our goal of generating high-quality fashion designs based on images of individuals. It is essential that our model incorporates various physical attributes of the person, such as complexion, body type, and shape, to generate clothing styles that are well-suited for the individual.

- ControlNet (7): The paper "Adding Conditional Control to Text-to-Image Diffusion Models" by Lvmin Zhang and Maneesh Agrawala introduces a new neural network structure called ControlNet that can be used to control the output of a diffusion model. The ControlNet takes an additional input condition, such as a depth map, a segmentation map, or a scribble, and uses it to guide the generation process. This solution can shows an exciting use case for fashion designers, who can now use this model to generate images in the required pose, which an replace the need for photoshoots. Although ControlNet offers a number of ways to customise the generated image through Stable Diffusion, it does not allow us to control the clothing which is generated. It generates clothes for simple prompts as can be seen in Fig 8.

(a) (b)

Figure 8: *Images generated by Stable diffusion model fine tuned on a fashion dataset. (a)Prompt: beautifully lit fashion portrait of black female marble statue with symmetrical face, the statue is wearing huge oversize quilted flowing floor length long puffer jacket by balenciaga, yeezy, y 3, yohji yamamoto, comme de garcon, rei kawakubo, drape, sharp focus, clear, detailed,, romantic, brutalist concrete architecture in the background, detailed, white, soft, symmetrical, vogue, editorial, fashion, magazine shoot, glossy (b) Prompt: A man with a symmetric face in a garden in jean overalls, white tshirt and a hat*

(a) (b) (c) (d)

Figure 9: *(a), (c) Input image for pose, (b), (d) Output image with prompt "a woman in a summer dress"*

4 Idea and Proposal

Our goal is as follows: Generate pose aware photo-realistic clothes for a target image, using an input image of the cloth texture, and a text prompt.

Generative AI in general, creates images in intractable manner. Previous work such as (4) try to constrain the generation using text. Extended work on (4) and (6) add both image and text as constraint. However, we discussed in good detail how our studies showed clear limitations to generate expected results.

By adding more constraints, we would be able to control the generation in certain directions and get images that are closer to what we expect. The idea goes against intuition that adding constraints helps in enhancing creativity which is shown in (8).

5 Our Work

Our idea for generating pose-aware photo-realistic clothes for a target image is based on several key factors that work together to produce high-quality results.

First, we recognize that certain information, such as cloth texture, cannot be fully conveyed in text and requires a visual context to be properly understood. Therefore, we incorporate both text and image inputs to our model, allowing us to generate clothing items that are both accurate and visually appealing.

We also use cloth segmentation to provide an image prior to the diffusion model and control regions that need to be in-painted. This helps to ensure that the clothing item is properly placed and scaled relative to the body in the target image.

In addition, we use augmentation to superimpose the cloth texture onto the segmentation mask. This allows us to generate a realistic-looking clothing item that accurately reflects the texture of the input image.

To further enhance the quality of our results, we use a stable diffusion model that takes into account the context of the prompt. For example, if the prompt is "Complete Woman's Dress in High Quality", we ensure that the inpainting process is focused on generating a woman's dress that meets the specified quality standards.

Finally, we recognize that when the texture image is superimposed onto the segmentation mask, the information about the body shape embedded in the dress can be lost. To address this, we use PIDM (Person Image Synthesis via Denoising Diffusion Model)(1) to generate images of the person wearing the generated clothing item in different poses, which we input as OpenPose landmarks. The architecture for the PIDM model is shown in Figure 11. By generating multiple images with different poses, we can better understand how the dress looks on the person and ensure that it is properly aligned with the body in the target image. This results in a more natural-looking and aesthetically pleasing final result.

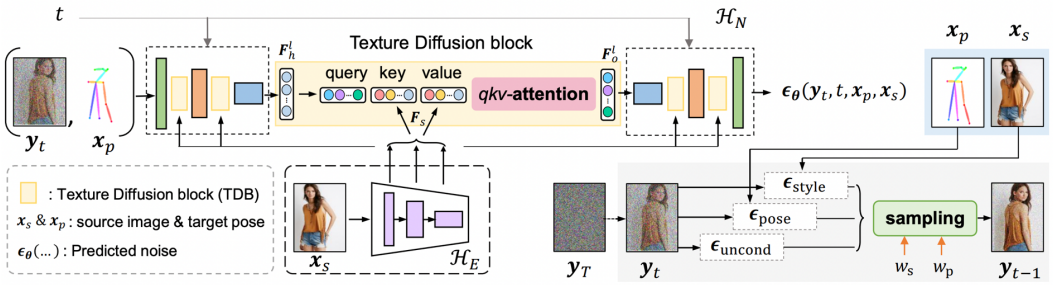


Figure 10: *PIDM Model Architecture (1)*

The flow chart for our implementation is shown in Figure 11 and our pipeline is as seen in Figure 12.

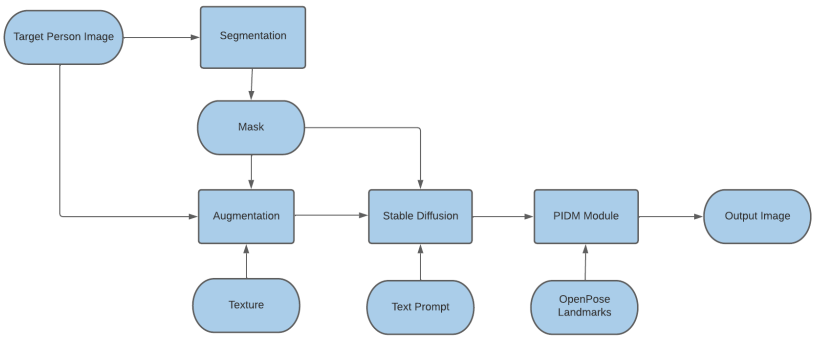


Figure 11: *Implementation Flow Chart*

Figure 12: *Pipeline Stages[L-R]: Input Image,Input Segmentation Mask,Input texture and color, Expected image result, OpenPose Landmark Input, Target image in the target pose*

6 Results

As discussed in our work, we were able to generate great results using our method described. The new clothes bear patterns that have very close resemblance to the texture provided. Additional detail to notice is that the reference model looks unchanged in terms of pose, face, and bodily features. We show results from using the two variations of Stable Diffusion models.

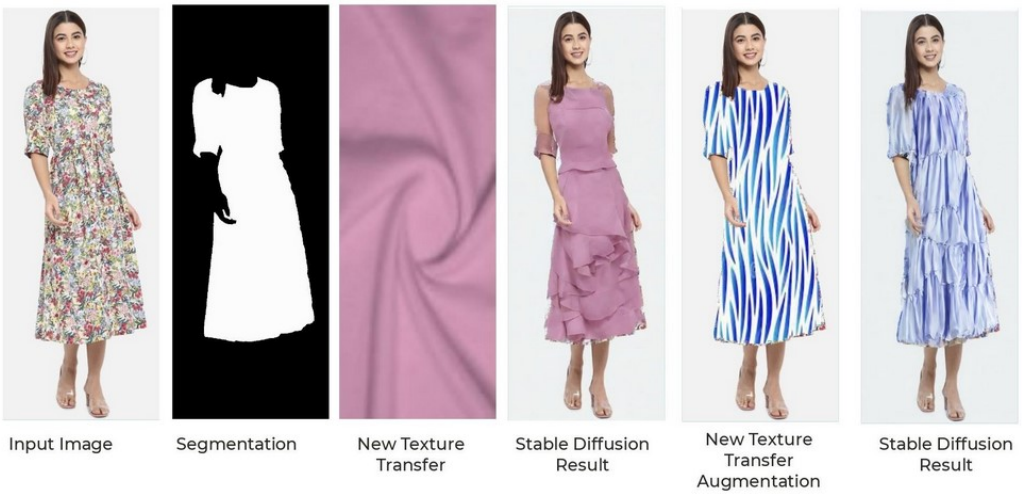


Figure 13: *Our results - Stable diffusion 2.1*

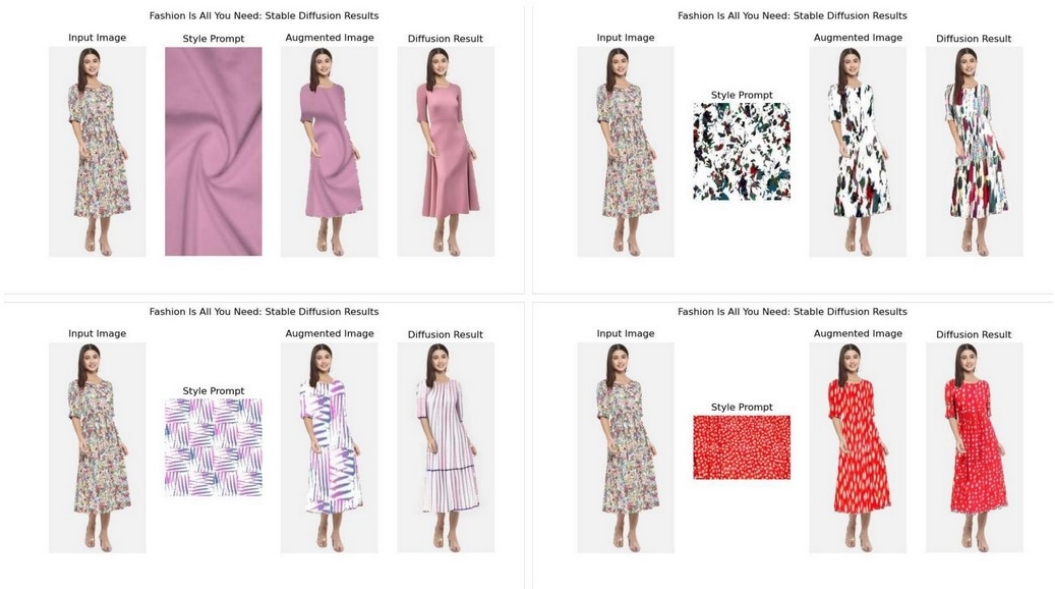


Figure 14: *Results generated with more textures and Stable diffusion 1.5*

It is important to note that the generated clothing items is heavily dependent on the accuracy of the segmentation of the clothes in the input image, as any errors in the segmentation could result in inaccurate and unrealistic clothing generation.

7 Future Work

1. While our current evaluation method involves visually inspecting the results generated by our model, we recognize the importance of having a well-defined quantitative metric for measuring the performance of our pipeline. A quantitative metric would allow us to compare different models and track improvements over time, as well as provide a more objective measure of performance. Moving forward, we plan to investigate and implement relevant metrics, in order to obtain a more comprehensive understanding of the accuracy and quality of our generated clothing items.

2. In the process of generating poses for models wearing new clothes, we have observed occasional facial distortions. To address this issue and achieve a more comprehensive solution, we are exploring the option of fine-tuning the PIDM model.

3. Currently, our pipeline involves feeding the results of one stage to the next manually. Moving forward, we plan to integrate our pipeline and automate the entire process. This would involve feeding in the input image of the person, texture image, OpenPose landmarks, and text prompt at the beginning of the pipeline, and receiving the final output of the pipeline at the end. The final output would be an image of the given person in the generated clothing item, in the poses defined by the input landmarks. By automating the end-to-end pipeline, we aim to increase efficiency, reduce errors, and ultimately produce more accurate and high-quality results.

8 Conclusion

We started out with an idea to generate clothes that are controlled by texture input. We experimented with multiple state-of-the-art generative models, but they did not produce the expected results. So we proposed a pipeline that takes advantage of segmentation, image processing techniques, and stable diffusion models to generate novel, fashionable, highly detailed clothes.

Our pipeline consists of the following steps:

Segmentation: We first segment the input image to identify the body and the clothes. Image processing: We then apply image processing techniques to the texture image to improve its quality. Stable diffusion model: We then use a stable diffusion model to generate the clothes on the body.

We evaluated our pipeline on a dataset of human images with different poses and clothing styles. We showed that our pipeline is able to generate novel, fashionable, highly detailed clothes that are consistent with the input texture and the body shape.

References

[1] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah, and F. S. Khan, "Person image synthesis via denoising diffusion model," 2023.

[2] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022.

[5] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," 2023.

[6] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," 2023.

[7] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[8] B. D. Rosso, "Creativity and constraints: Exploring the role of constraints in the creative processes of research and development teams," *Organization Studies*, vol. 35(4), p. 551–585, 2014.