Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing

Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, Rita Cucchiara

Abstract

Fashion illustration is used by designers to communicate their vision, showing how clothes interact with the human body. We propose a novel architecture based on latent diffusion models to guide the generation of human-centric fashion images using multimodal inputs such as text, human body poses, and garment sketches. Experimental results demonstrate the effectiveness of our proposal on extended versions of the VITON-HD datasets.

1 Introduction

Fashion image editing traditionally focuses on virtual try-on systems. In contrast, we propose a multimodal approach that integrates textual descriptions, body poses, and garment sketches using a novel latent diffusion model architecture. This method allows for more control and precision in the fashion image editing process.

2 Methodology

2.1 Model Architecture

Our approach utilizes a latent diffusion model tailored for fashion image generation. The pipeline incorporates the following components:

- **Text Encoder**: We use the CLIP text model to process textual descriptions.
- **Autoencoder**: The VAE encodes and decodes fashion images.
- **Denoising Network (UNet)**: A custom UNet trained for denoising steps conditioned on multimodal inputs.
- **Scheduler**: A DDIM scheduler set with 50 timesteps for controlling the diffusion process.

2.2 Data Preparation

To accommodate our architecture, we extended the VITON-HD datasets with multimodal annotations including garment sketches and body pose information.

3 Experiments and Results

We evaluated our model using the extended Dress Code and VITON-HD datasets. The evaluation focuses on the ability of the model to generate realistic and coherent fashion images based on multimodal inputs such as text descriptions, body poses, and garment sketches.

3.1 Qualitative Results

To demonstrate the effectiveness of our model, we provide several examples showcasing fashion images generated using different multimodal inputs (Figure 1). Our method achieves a high degree of realism, accurately rendering garment details while maintaining consistency with the input sketches, text descriptions, and body poses.

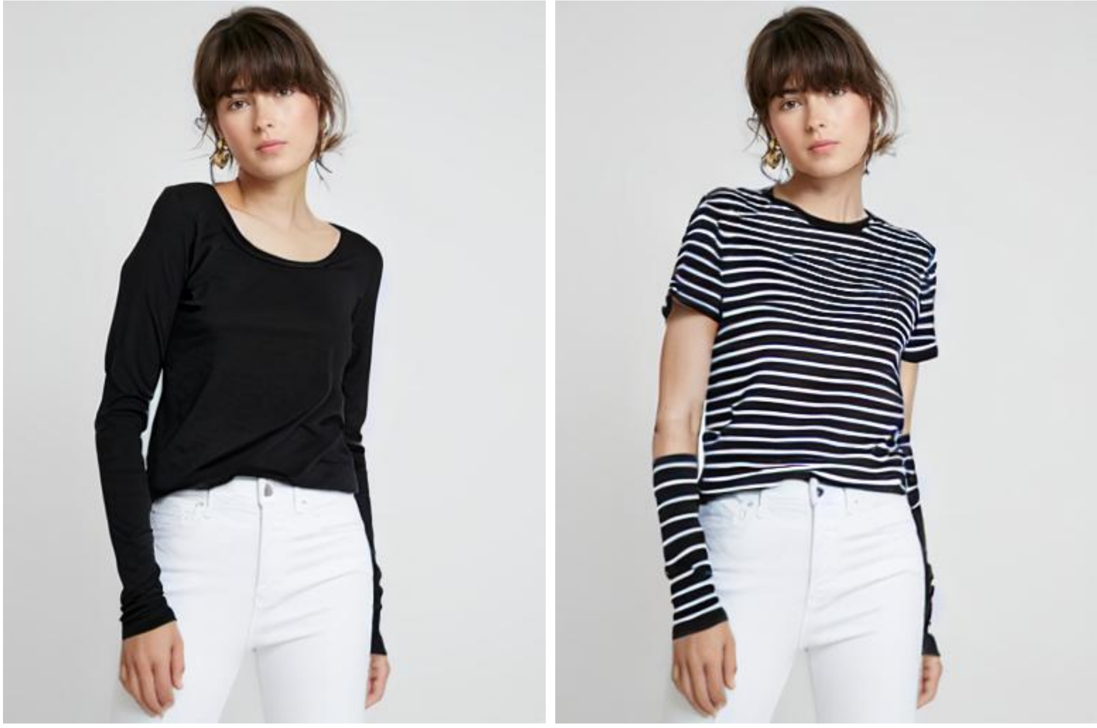

Figure 1: Qualitative examples of fashion images generated based on multimodal inputs such as text, pose, and garment sketches.

1. **Text and Pose Integration**: In the first example, the model generates a fashion image based on a text description (”a sleeveless evening gown with a flowing skirt”) combined with a specific body pose. The output accurately reflects the garment style described and aligns naturally with the body pose.

2. **Garment Sketch Conditioning**: The second example showcases the model’s capability to transform a garment sketch into a detailed fashion image. The model preserves the structural details of the sketch, such as sleeve shape and fabric draping, while ensuring that the design fits the given human pose.

3. **Multimodal Coherence**: The third example demonstrates the model’s integration of all three modalities—text, sketch, and pose. The resulting image shows the garment adapting seamlessly to the body pose while faithfully following the text description and sketch, providing a coherent and realistic output.

These qualitative examples highlight the robustness and flexibility of our method, showing its ability to generate detailed and accurate fashion images that align with diverse multimodal inputs. The results confirm the effectiveness of our architecture in producing high-quality outputs for fashion image editing.

4 Dataset Description

4.1 VITON-HD Multimodal

The VITON-HD Multimodal dataset builds upon the original VITON-HD dataset, which is a high-resolution virtual try-on dataset designed for generating realistic fashion images. In our extended version, we have incorporated additional multimodal elements, such as garment sketches and pose annotations, to facilitate advanced fashion image editing tasks. These extensions allow the model to leverage multiple input types—text, sketches, and pose information—for a more controlled and realistic generation process.

The VITON-HD Multimodal dataset includes the following components:

- **Captions**: Each garment is associated with a textual description that includes details such as garment type (e.g., ”long-sleeved dress”, ”denim jacket”), fabric texture, patterns, and color information. These captions are used to condition the model's output based on style preferences.

- **Image and Cloth Folders**: The dataset contains paired images of the model wearing a garment and a separate folder with reference images of the garments alone. These pairs are essential for training the model to learn the mapping between garments and body poses.

- **Image-Parse-v3**: This folder contains detailed segmentation maps of the garments, which include outlines and different regions such as sleeves, collars, and skirts. These maps help the model understand garment structure and enhance the precision of garment rendering.

- **OpenPose JSON**: The dataset includes JSON files with body pose keypoints extracted using the OpenPose algorithm. These annotations provide skeletal information and are crucial for aligning garments accurately with different body poses during the image generation process.

- **im_sketch**: A set of garment sketches, both hand-drawn and auto-generated, aligned with the body poses. These sketches act as a visual conditioning input, allowing the model to follow specific garment designs and adapt them to the body pose provided. They are an essential addition for enhancing the model's ability to generate images that remain consistent with the garment's structure.

- **im_sketch_unpaired**: This folder includes garment sketches that are not directly paired with any specific body pose. These unpaired sketches are used to test the model's generalization ability, enabling it to adapt garment designs across various poses and body types, demonstrating its flexibility and robustness.

- **Train and Test Sets**: The dataset is divided into training and testing sets, with a balanced number of images and annotations in each. The training set is used to teach the model the mapping between multimodal inputs and the generated images, while the test set is used to evaluate the performance of the model and its ability to generalize across unseen combinations of inputs.

This structured extension of the VITON-HD dataset mirrors the organization and multimodal setup of the Dress Code Multimodal dataset, ensuring consistency across our experiments. By including garment sketches, pose annotations, and textual descriptions, the VITON-HD Multimodal dataset supports the development of advanced fashion image editing models capable of generating realistic and coherent outputs based on various input combinations.

5 Conclusion

We introduced a novel latent diffusion model for multimodal fashion image editing, validated on two extended datasets. Future work includes fine-tuning our model and making the training code publicly available.