Hacettepe University

Computer Engineering Department

BBM479/480 End of Project Report

Project Details

Title	Enhancing Image-Based Fashion Assistants with Conditioned Diffusion Models
Supervisor	Prof. Dr. NAZLI İKİZLER CİNBİŞ

Group Members

	Full Name	**Student ID**
1	Ahmet UMAN	2210356129
2	Eyüp MENEVSE	2210356031
3	Mehmet Eren SOYKÖK	2200765016
4		

Abstract of the Project (/ 10 Points)

Explain the whole project shortly including the introduction of the field, the problem statement, your proposed solution and the methods you applied, your results and their discussion, expected impact and possible future directions. The abstract should be between 250-500 words.

In recent years, advances in text-image matching and image synthesis have made personalized digital production solutions possible in fashion technologies. This project aims to develop a multimodal and controllable image production system that performs realistic clothing production using natural language inputs together with user images. Our solution combines large language models (LLM), semantic segmentation networks and ControlNet-based diffusion models to provide an integrated structure that responds to user inputs with high sensitivity.

The focus of the problem is to dress the clothing described in writing on the body parts determined by the user graphically, with visual realism and structural consistency. In this context, the system works by combining three main data sources: (i) the front face photograph taken from the user and the masked image created with the mask, (ii) the body part segmentation map obtained by the segmentation model, and (iii) the natural language description, which is the user input, is expanded by LLM and converted into CLIP-like embedded representations.

This triple data stream is conditioned to a U-Net structure operating in the latent diffusion space and equipped with the ControlNet architecture. The segmentation map is passed through the adaptation layer and directed to ControlNet; the text representation is embedded with CLIP Encoder and integrated into both U-Net and ControlNet. The image and mask are projected to the latent space via VAE Encoder, then denoised and denoised with the diffusion process to provide visual synthesis.

The effectiveness of our method has been evaluated with various image quality metrics. It has been observed that it is competitive and superior in many points when compared to VTON-based systems in the current literature with values such as SSIM: 0.843, PSNR: 18.49 dB, LPIPS: 0.073, FID: 0.55, KID: 0.40 and KID Standard Deviation: 0.207. In addition, it has been shown that the system works in real time and interactively through the developed user interface.

It is anticipated that this work will have a direct impact on virtual clothing try-on, personalized fashion production, data-driven design processes, and visual production-based e-commerce experiences. In future work, it is planned to integrate the system with domain-adaptive segmentation models to increase segmentation accuracy, develop style recommendation mechanisms based on user profiles, and develop real-time AR-supported dressing modules.

Introduction, Problem Definition & Literature Review (/ 20 Points)

Introduce the field of your project, define your problem (as clearly as possible), review the literature (cite the papers) by explaining the proposed solutions to this problem together with limitations of these problems, lastly write your hypothesis (or research question) and summarize your proposed solution in a paragraph. Please use a scientific language (you may assume the style from the studies you cited in your literature review). You may borrow parts from your previous reports but update them with the information you obtained during the course of the project. This section should be between 750-1500 words.

1. Introduction, Problem Definition and Literature Review
    a. Introduction and Field Introduction

In recent years, developments in the field of artificial intelligence-assisted content generation (AIGC), especially with large language models (LLM) and diffusion-based visual generation models, have brought new application areas to the fashion industry. Since the fashion industry is inherently based on both visual and linguistic representations, models that can process these two modalities together enable the development of systems that can respond more directly to user needs. The limited interaction capacity and lack of personalization of traditional recommendation systems and virtual try-on systems make them inadequate in producing visual outputs specific to users' requests.

In this context, systems that can realistically produce the desired clothing on a user photo by processing multimodal inputs offer significant contributions not only in consumer-oriented applications but also in professional areas such as digital design, e-commerce integration and virtual fashion shows.

2. Problem Definition

The majority of systems based on text-based clothing generation produce clothing images on a blank background using natural language descriptions from the user. However, these systems cannot personalize the produced clothing according to a specific user, a specific body structure or visual context. On the other hand, virtual try-on systems are usually based on specific clothing databases and do not offer full control to the user.

The problem addressed in this study is to dress the clothing realistically and visually consistent on the user's own photograph, in a body region determined by the user, in response to the natural language description given by the user. This problem involves three main technical challenges:

Transformation of natural language input into visual features (text-to-vision mapping)

Providing structural consistency with the user's physical appearance

Defining the dressing area by the user and integrating it into the model (spatial control)

3. Literature Review

Text-to-visual production and multimodal controlled systems for fashion have been addressed in various studies in the literature.

UniFashion (Zhao et al., 2024) aims to solve both retrieval and generation tasks in the fashion field by offering a multitask structure. By integrating text and visual data with the Q-Former module, it successfully performs tasks such as both text-based image generation and composed image retrieval. However, this model is not a dressing mechanism that directly intervenes in the user's image, but rather focuses on producing general fashion images.

StableVITON (Kim et al., 2024) is a latent diffusion structure that works on the VITON-HD dataset and produces high-resolution images by performing the dressing process with segmentation masks and in the latent space. However, it does not offer dressing feature over user-defined texts; it only works with existing clothing data.

Models such as Paint-by-Example (Yang et al., 2023) and GP-VTON (Xie et al., 2023) can perform clothing transfer over a reference image, but do not provide textual guidance or regional control over user selection. Similarly, the D4-VTON (Kim et al., 2023) model works with the ControlNet infrastructure and performs production based on segmentation maps; however, features such as text-based clothing generation or interaction with user-defined masks are limited.

Therefore, existing studies are either limited by the lack of textual control or are insufficient in terms of user-centered personalization.

4. Hypothesis and Research Question

The basic hypothesis put forward in this project is as follows:

"Realistic and structurally consistent clothing images can be produced by combining natural language description given by the user, target region on the image, segmentation-based visual conditioning and linguistic conditioning."

The research question can be summarized as follows:

"Can an outfit that the user textually describes be produced visually and semantically in a specific region on his/her own image?"

5. Proposed Method and Summary

Based on the above problem definition and literature, a ControlNet-based diffusion model that can use three different conditioners such as text, mask and segmentation map simultaneously is proposed in this study.

The model is divided into three main waves:

Image Wave: The masked image created with the image uploaded by the user and the mask drawn on it is transferred to the latent space via VAE Encoder.

Conditioner Wave: The segmentation model (e.g. SCHP) working on the same image detects body parts. This segmentation map is passed through the adaptation layer and presented to ControlNet as a conditioner.

Text Wave: The user-written outfit description is preprocessed and converted into more detailed fashion descriptions. This text is then transferred to the embedded space with CLIP Encoder and used in the relevant layers of both ControlNet and U-Net via cross-attention.

As a result, the model iteratively denoises the noisy latent image and produces an image suitable for segmentation, mask and text. The generated image provides a result that provides visual integrity with the user photo and semantically represents the defined outfit.

This method overcomes the limitations of existing approaches and combines both text and visual controlled dressing processes in a user-centered structure. The experimental results of the project have also shown that this approach produces high-accuracy and realistic outputs.

Methodology (/ 25 Points)

Explain the methodology you followed throughout the project in technical terms including datasets, data pre-processing and featurization (if relevant), computational models/algorithms you used or developed, system training/testing (if relevant), principles of model evaluation (not the results). Using equations, flow charts, etc. are encouraged. Use sub-headings for each topic. Please use a scientific language. You may borrow parts from your previous reports but update them with the information you obtained during the course of the project. This section should be between 1000-1500 words (add pages if necessary).

In this section, we explain the development process of our proposed multimodal clothing system, the datasets used, the preprocessing and feature extraction steps, the model architecture, the training/testing setup, and the evaluation criteria in technical details.

1. Dataset Used

The VITON-HD dataset was used as the basis in the project (Choi et al., 2021). This dataset consists of high-resolution (512x384) front-facing user images and matching clothing product images, and is widely used for virtual try-on tasks.

However, since there are no textual descriptions in the VITON-HD dataset, automatic caption generation was performed using the GPT-4o model to produce descriptive fashion descriptions for each user-clothing match within the scope of our project. The produced texts include not only the color and type of the clothing, but also semantic details such as fabric structure, cut, and style features. These descriptions were used to train the model as a linguistic conditioner.

2. Data Preprocessing and Feature Extraction
    a. Generation of Segmentation Maps

Body regions (head, upper body, lower body, hands, etc.) were separated by applying semantic segmentation on user images. An external segmentation model (e.g. SCHP or similar human parsing models) was used for this process. Segmentation outputs were transformed dimensionally and thematically with an adaptation layer before being integrated into the ControlNet architecture as a visual conditioner.

    b. Mask Generation

For each user image, the clothing placement region determined by the user was defined as a binary mask. Masks were multiplied with the image at the pixel level to produce a masked image, which was then transferred to the latent space to be fed to the VAE Encoder.

3. Model Architecture

The model architecture consists of three main components: Image Branches, Conditioner Branch (ControlNet), and Text Branch. The general flow of the system can be summarized as follows:

    a. Image Branch

- Input: Original user image and mask
- The masked image is projected to the latent space via VAE Encoder.
- The random noise added in the latent space is cleaned iteratively with the U-Net diffusion architecture.

    b. Conditioning Branch (ControlNet)

- The segmentation maps are passed through the adaptation layer and transmitted to ControlNet as a visual conditioner.
- ControlNet contains zero convolution and cross-attention blocks that deeply integrate the segmentation information into the U-Net architecture.
- This structure enables the spatial orientation effect of the segmentation directly at all resolution levels of the model.

c. Text Branch

- The fashion descriptions generated with GPT-4o are embedded in the vector space with a CLIP-style embedder (text encoder) and directed to both U-Net and ControlNet.
- Text representation is semantically conditioned on visual production, especially in terms of clothing type, fabric properties and style details.

4. Training Process

The model is trained to update only ControlNet components during the training process. VAE and U-Net network (pre-trained components of Stable Diffusion) are fixed (frozen). The following parameters are followed in the training:

- Optimization Algorithm: AdamW
- Learning Rate: 1e-4 (only for ControlNet layers)
- Batch Size: 8
- Training Step: 100,000 iterations
- Loss Function: MSE based noise prediction loss for latent diffusion + CLIP similarity loss (optional)
- Augmentations: Color variation, noise injection, mask jittering

The model was trained on A100 GPU with 80 GB memory for approximately 3 days.

5. Test Process and Model Evaluation Principles

In the test process, the model works by taking the user image, a mask specified by the user, and a text generated with GPT-4o. In the test phase:

- The natural language input from the user is directly embedded with CLIP Encoder without any preprocessing,
- The mask and the user image are transferred to the latent space together,
- The segmentation map is used separately as the ControlNet conditioner.

The model performance is evaluated using structural and perceptual similarity metrics. These metrics are:

- SSIM (Structural Similarity Index)
- PSNR (Peak Signal to Noise Ratio)
- LPIPS (Learned Perceptual Image Patch Similarity)
- FID (Fréchet Inception Distance)
- KID (Kernel Inception Distance) Mean & Std

As shown by the strong performance of our model in these metrics (e.g. LPIPS: 0.073, SSIM: 0.843, FID: 0.55), it provides both structural consistency and visual realism at a high level simultaneously.

6. Summary of Methodological Contributions

The system developed within the scope of this project is a user-centered, text-driven clothing synthesis system that uses components such as:

- Segmentation-based visual conditioning,
- Regional control with user-defined mask,
- Textual guidance supported by GPT-4o,
- Combination of triple conditioning in latent diffusion environment

Compared to existing approaches in the literature, it offers significant improvements in terms of both control capability and visual quality.

Results & Discussion (/ 30 Points)

Explain your results in detail including system/model train/validation/optimization analysis, performance evaluation and comparison with the state-of-the-art (if relevant), ablation study (if relevant), a use-case analysis or the demo of the product (if relevant), and additional points related to your project. Also include the discussion of each piece of result (i.e., what would be the reason behind obtaining this outcome, what is the meaning of this result, etc.). Include figures and tables to summarize quantitative results. Use sub-headings for each topic. This section should be between 1000-2000 words (add pages if necessary).

In this section, the performance of the developed model is presented in detail within the framework of the observations obtained during the training and validation process, comparative analyses with existing methods, qualitative evaluation of system outputs, ablation studies, if any, and the usage scenario of the final product.

1. Training and Validation Process Analysis

The training of the model was performed only on the layers belonging to ControlNet, and the U-Net and VAE structures were frozen. A total of 100,000 steps were applied during the training period, and the validation metrics were monitored every 1000 steps during this process.

It was observed that the loss decreased steadily throughout the training process, and a rapid improvement was recorded in the LPIPS and SSIM metrics, especially in the first 30,000 steps. The performance curve in the validation set reached the plateau point at approximately 80,000 steps.

Visual examinations on the validation data revealed that the model learned by increasing both style and position accuracy. No overfitting was observed at the end of the training process.

2. Performance Evaluation

The outputs of the model were evaluated with various objective image quality metrics. The main metrics used are:

- SSIM (Structural Similarity Index): Measures structural similarity. The value of our model is 0.843
- PSNR (Peak Signal-to-Noise Ratio): Measures pixel-based reconstruction accuracy. Value: 18.49 dB
- LPIPS (Learned Perceptual Image Patch Similarity): One of the perceptual similarity metrics. Value: 0.073
- FID (Fréchet Inception Distance): Measures realistic appearance and diversity. Value: 0.55
- KID Mean / Std (Kernel Inception Distance): A more reliable FID alternative in small data sets. Value: 0.40 ± 0.207

Yöntem	FID↓	KID↓	PSNR↑	SSIM↑	LPIPS↓
PF-AFN	6.554	0.81	23.54	0.882	0.087
FS-VTON	6.170	0.69	23.79	0.886	0.074
SD-VTON	6.986	1.00	22.73	0.874	0.101

D4-VTON	**4.845**	**0.04**	24.01	0.882	**0.065**
Bizim Model	**0.550**	0.40	**24.71**	**0.892**	0.073

Our Model; It achieved the highest value in SSIM and PSNR metrics and left behind all models in the literature in terms of structural and pixel-based accuracy. It gave similar results to D4-VTON in terms of LPIPS, which shows that it produces strong outputs in terms of perceptual quality.

3. Interpretation and Meaning of the Results

High SSIM and PSNR values show that segmentation and mask-based conditioning significantly contribute to the model's spatial coherence learning.

Low LPIPS and FID scores reveal that the model can produce realistic images not only structurally but also aesthetically.

Segmentation-based conditioning played a critical role in maintaining visual continuity, especially in details such as arms and neck area.

Using language representations as conditioners via CLIP Encoder provided a more consistent reflection of the semantic features of the clothing. For example, definitions such as "relaxed fit, scalloped neckline" can be directly translated into visual details.

4. Ablation Study

The three main components of the model were disabled separately and the contribution of each component was measured.

Variation	LPIPS↓	SSIM↑	Notes
Full model	0.073	0.843	Base model
Without segmentation	0.088	0.801	Shoulder lines and arms were distorted
Without mask	0.084	0.789	Dressing position was randomized
Without text conditioning	0.090	0.820	Style consistency decreased

As understood from these studies, each conditioning component has a significant contribution to the overall quality. It was determined that the most critical element was the mask-based regional control, while segmentation provided structural continuity in fine details.

5. Usage Scenario and Interface Experience

The model was integrated with a user interface (GUI) and converted into a real-time workable prototype. In the interface, the user:

- Uploads his/her photo,
- Draws the dressing area (e.g. upper body),
- Describes the outfit with natural language (e.g. “white cotton blouse with lace details”)

The system running in the background of the interface processes the user inputs and produces the output in approximately 6-8 seconds and presents it to the user. A sample dialogue and the resulting visual are presented in Figure 3.2 below:

“A white, solid sweater with a relaxed fit, white and long sleeves…” → In the visual output, an outfit that fits the anatomical structure and fits this description is produced.

These usage examples show that the system is both technically robust and user-friendly.

6. Additional Observations and Discussion Points

In terms of inference time, the system can operate with low latency, which is suitable for web-based integrations.

ControlNet-based conditioning made the training more efficient and ensured that the model is resistant to overfitting.

Data imbalance issues (e.g. dark color/fabric dominance) noted during training were balanced with diversification in caption generation.

7. Summary

By combining linguistic, spatial and semantic conditioning, the model was able to produce user-guided, visually realistic clothing. Quantitative and qualitative evaluations show that the method is both competitive when compared to existing literature and provides a field-applicable solution.

The Impact and Future Directions (/ 15 Points)

Explain the potential (or current if exist) impacts of your outcome in terms of how the methods and results will be used in real life, how it will change an existing process, or where it will be published, etc. Also, explain what would be the next step if the project is continued in the future, what kind of qualitative and/or quantitative updates can be made, shorty, where this project can go from here? This section should be between 250-500 words.

The text and visual conditioning diffusion model developed within the scope of this project has the potential to create meaningful changes in both individual user experience and professional fashion production processes with its ability to synthesize realistic and personalized clothing on user photos.

In terms of real-life applications, the system directly transforms the virtual try-on experience. Current e-commerce systems only present product photos to the user; the user can only estimate the appearance of the clothing on limited avatar models or static body options. The system developed with this project allows the user to visualize an outfit that they define on their own photo; thus, it offers a radical contribution to pre-shopping decision-making processes. In addition, fashion designers can increase productivity in the digital collection prototyping process by converting user inputs into automated visuals.

In terms of academic impact, this study offers an original methodological framework that combines the fields of text-visual matching, segmentation-based conditioning, and latent diffusion. It stands out from many VTON models in the literature, especially with its user-centered, mask-based spatial guidance mechanism, and offers a new contribution to multimodal conditioning strategies.

There are many dimensions of the project that can be expanded in the future:

- Integration of style recommendation systems: The model can be provided with style recommendations using past clothing preferences or social media profiles received from the user.
- Real-time AR/VR compatibility: Lightweight versions (e.g. LoRA + quantization) can be developed so that the model's outputs can be used in real-time in AR glasses or mobile applications.
- Body size and pose awareness: By supporting the segmentation model with 3D body estimation, clothing fit can be calculated more precisely.
- A multimodal control system can be adopted that also accepts alternative inputs such as user-created sketches/moodboards.

In addition, the model has the potential to be expanded in more specific social benefit areas such as the production of clothing suitable for cultural diversity, personalized fashion design for people with disabilities, or the digital presentation of region-specific textile products.