IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"OOTDiffusion: Outfitting Fusion Based Latent Diffusion For Controllable Virtual Try-On"

[1]Miss. Farah Subhedar, [2]Miss. Madhura Nigavekar, [3]Miss. Sanika Lohar, [4]Miss. Rutuja Patil, [5]Prof. Rabiya Kothiwale

[1, 2, 3, 4]Student, [5]Assistant Professor

Department of CSE (Data Science),

D. Y. Patil College of Engineering & Technology, Kolhapur, India.

Abstract: OOTDiffusion is a groundbreaking method for virtual try-on that delivers exceptional realism and control. By leveraging the power of latent diffusion models and incorporating innovative techniques like outfitting fusion and dropout, OOTDiffusion generates incredibly natural and lifelike images. It allows for precise customization of garment details and their placement on the body, providing a highly personalized virtual try-on experience. This method is efficient, versatile, and capable of handling a wide range of garment types and body shapes. OOTDiffusion could change online shopping by giving customers a more engaging and enjoyable way to try on clothes virtually. It also helps fashion retailers by lowering return rates and improving customer satisfaction. We introduce OOTDiffusion, a new system for realistic and customizable virtual try-ons. Unlike other methods, it accurately matches clothing to the person's body without unnecessary adjustments. We also add a feature called outfitting dropout during training, allowing us to control how prominent the clothing details are. Our tests on popular datasets show that OOTDiffusion produces high-quality try-on images for various people and outfits. It outperforms other methods in both realism and control, marking a significant advancement in virtual try-on technology.

Keywords – OOTDiffusion, Virtual Try-on, Realism, Control, Diffusion Models, Garment Customization, Body Fit, Customer Satisfaction

I. INTRODUCTION

Image-based virtual try-on (VTON) is an exciting technology in e-commerce that enhances the shopping experience for customers and lowers advertising costs for clothing retailers. VTON aims to create an image of a person wearing a specific piece of clothing. However, there are two main challenges. First, the images need to look realistic to avoid looking awkward. Many recent methods use generative adversarial networks (GANs) or latent diffusion models (LDMs) for this purpose. Older GAN methods often struggle with creating realistic clothing folds, lighting, and human bodies. That's why newer approaches are focusing on LDMs, which do a better job at making the images look more lifelike.

Need of the Work: Online shoppers want more interactive and realistic ways to try on clothes. Virtual try-on systems help with this by showing how clothes would look without trying them on physically. This improves customer satisfaction and reduces product returns. But making these images look real and detailed is still a challenge.

Existing Systems: Older virtual try-on methods use GANs and image warping. While they started the trend, they often struggle with making clothes look natural, keeping textures intact, and fitting clothes to different body shapes. Even with newer techniques like latent diffusion models, fine clothing details often get lost.

Proposed System: OOTDiffusion solves these problems using advanced diffusion models and better customization tools. It keeps clothing details like texture, text, and patterns clear, and works well with different body types and poses. This makes virtual try-on more realistic, flexible, and useful for both customers and fashion retailers.

This proposed system aims to advance the current state of virtual try-on technologies by providing a more realistic, flexible, and user-friendly experience. With OOTDiffusion, users can easily customize and experiment with different clothing combinations, ensuring that the final rendered images are both visually appealing and practically useful for decision-making. The model's ability to handle a wide range of poses and body types further extends its applicability across various user demographics, making it a valuable tool for both consumers and retailers in the fashion industry.

II. Literature Survey

Yuhao Xu and colleagues' paper, OOTDiffusion, presents a novel VTON method using pretrained latent diffusion models and outfitting UNet to align garments with the target body. This approach, featuring outfitting dropout for adjustable features, shows improved realism and controllability, as tested on VITON-HD and Dress Code datasets. The work was submitted on March 4, 2024, and revised on March 7, 2024. OOTDiffusion faces issues with cross-category try-ons and detail alteration, indicating a need for better datasets and processing methods.

The study by Gou et al. (2023) looks at using a diffusion model to create high-quality virtual try-ons, focusing on accurately transferring clothing details between images. While diffusion models show promise compared to traditional GAN-based methods, challenges remain, particularly with small and intricate patterns. The method struggles with precise details, such as small writing on clothing, due to potential loss in the latent space during the inpainting process. Despite these issues, the approach generally provides consistent result for less detailed patterns.

The study by Morelli et al. (2023) introduces LaDI-VTON, a virtual try-on system leveraging latent diffusion models to create realistic images of garments on target models. While the approach effectively captures the overall shape of logos and textual patterns on garments, it struggles with accurately rendering readable and coherent text. This limitation is attributed to the constraints of Stable Diffusion-based architectures, which may be improved with a non-latent diffusion approach, albeit at a higher computational cost.

The 2022 CVPR paper by Robin Rombach and colleagues introduces Latent Diffusion Models (LDMs). These models improve the creation of high-quality images by using diffusion techniques in a simplified version of the image data, which saves on processing power. LDMs include cross-attention layers that allow for flexible input options, like text and bounding boxes, and they perform well in tasks like filling in missing parts of images and generating images based on categories. However, LDMs can be slower at generating images compared to GANs and may struggle with tasks that need very precise results.

The paper by Ge et al. (2021) introduces the Disentangled Cycle-consistency Try-On Network (DCTON) to address challenges in virtual try-on, where the goal is to replace clothes on a person with desired garments from a shop. Traditional methods, which use in-painting or cycle consistency, fail to differentiate between clothing and non-clothing areas, affecting the quality of tryons. DCTON improves upon these methods by disentangling critical components such as clothes warping, skin synthesis, and image composition. This approach enables highly-realistic try-on images and can be trained in a self-supervised manner through cycle.

III. Methodology

1. Text Input Collection

User provides a text prompt describing the desired clothing and appearance.

Example: “A red floral dress on a woman in a standing pose.”

2. Data Preprocessing

The text input is tokenized and normalized for compatibility with the model pipeline.

Ensures uniform and structured input for downstream processing.

3. Feature Extraction

Semantic features are extracted from the preprocessed text.

These features capture meaning and intent from the user input.

4. Encoding with CLIP

CLIP Text Encoder processes the text features.

CLIP Image Encoder processes image features (if reference image is used).

Both encoders embed data into a shared latent space, enabling alignment of visual and textual data.

5. Outfitting Fusion

Combines encoded text and image features into a unified latent representation.
Enables precise alignment of garment descriptions with visual outputs.
Allows integration of multiple clothing elements into one cohesive image.

6. Latent Diffusion Image Generation

A latent diffusion model uses the fused features to generate realistic images.
Ensures natural lighting, body fit, and high-quality textures of garments.

7. Outfitting Dropout

Introduced during model training to control the visibility and influence of clothing features.
Enables flexibility in emphasizing or de-emphasizing certain garment details.

8. Image Refinement

Generated images may be enhanced using techniques like style transfer or image editing.
Improves realism and visual appeal.

9. Output Generation

The final, high-quality, realistic try-on image is produced and displayed to the user.

10. User Interaction

Interface built using Gradio, allowing users to upload inputs and view results.
Backend is managed by Flask, handling data processing and communication.

IV. SYSTEM DESIGN & IMPLEMENTATION

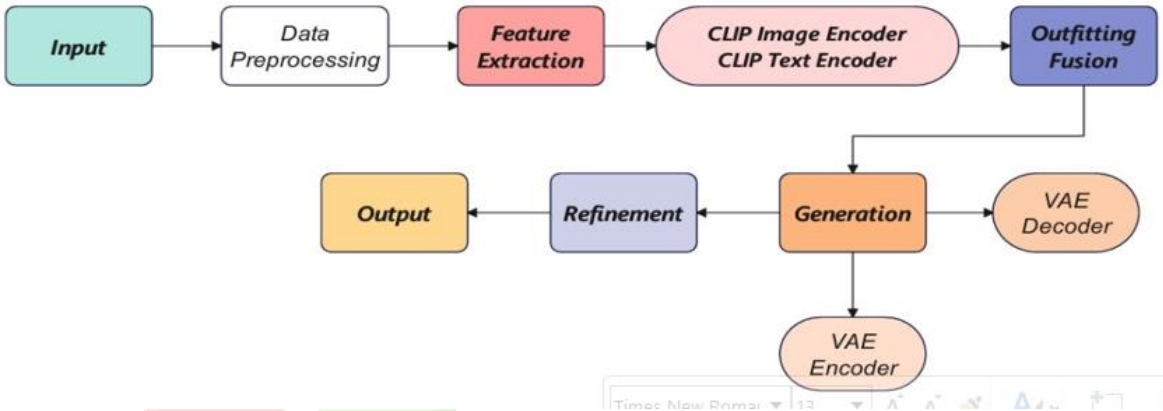


Fig. System architecture of "OOTDiffusion: Outfitting Fusion Based Latent Diffusion for Controllable Virtual Try-on"

1. Input: The process starts with text input, which describes the desired image.
2. Data Preprocessing: The text input is preprocessed to prepare it for further processing. This might involve tasks like tokenization or normalization.
3. Feature Extraction: Features are extracted from the preprocessed text input. These features represent the semantic meaning of the text.
4. CLIP Image Encoder and CLIP Text Encoder: The extracted features are encoded using two CLIP models: one for images and one for text. These models learn to represent images and text in a common latent space.
5. Outfitting Fusion: The encoded image and text features are fused to combine information from both modalities. This fusion step helps to align the text description with the generated image.
6. Generation: A generative model, such as a VAE (Variational Autoencoder), is used to generate an image based on the fused features. The VAE decoder generates the image from a latent representation.
7. Refinement: The generated image may be refined through further processing, such as image editing or style transfer.
8. Output: The final generated image is the output of the process.

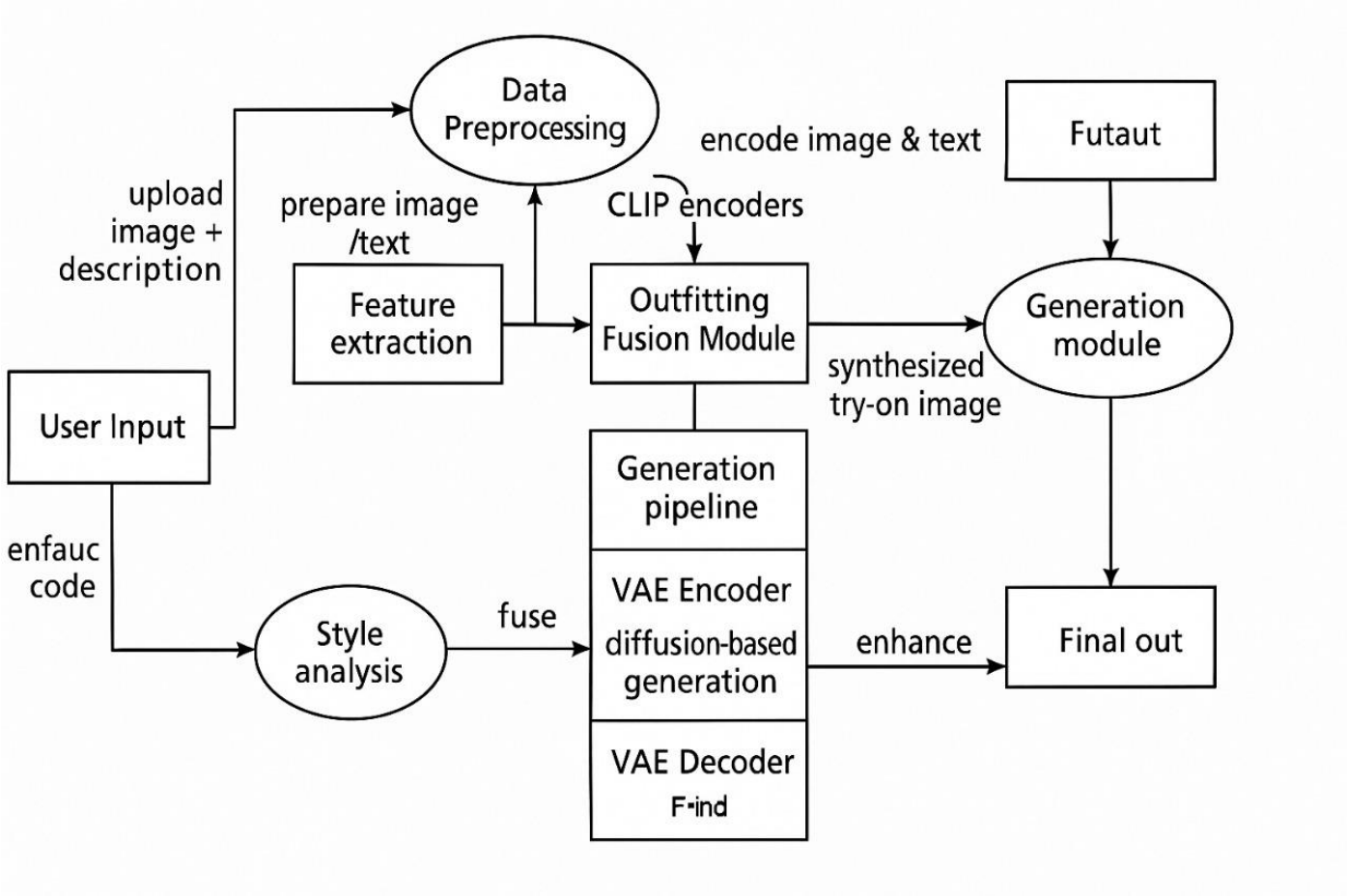


Fig. Implementation Diagram

Implementation Details:

The virtual try-on system, called OOTDiffusion, works by taking a person's image and the name of the clothing they want to try on like "red jacket" and using AI to show how it would look on them. First, the user uploads a photo and types in the clothing description using an easy-to-use web interface made with a tool called Gradio. When the user clicks submit, the information is sent to a server built using Flask, which handles all the behind-the-scenes work.

On the server, the clothing description (text) and the uploaded image are prepared and processed. The system uses a special AI tool called CLIP, which understands both images and text. It turns both the photo and the clothing description into numbers (features) that the computer can understand. These two sets of information are then combined in a process called Outfitting Fusion, which helps the system match the clothing properly to the person's body.

After combining the details, the system uses a powerful image-generating AI called a Latent Diffusion Model (LDM) to create a realistic image of the person wearing the clothes. To make the results more flexible, a special method called Outfitting Dropout is used during training to control how clearly details like text, patterns, or logos appear on the clothes.

Once the image is created, it's cleaned up and refined using tools like OpenCV, and the final picture is sent back to the user on the website. Other tools like NumPy help with calculations, random helps make different results if needed, and dotenv keeps sensitive info like passwords and API keys safe.

Algorithm:

1. User Input:

The user uploads a photo and types a description of the clothing (e.g., "blue t-shirt") using the Gradio web interface.

2. Request Handling:

The input is sent to the backend server built with Flask.

3. Data Preprocessing:

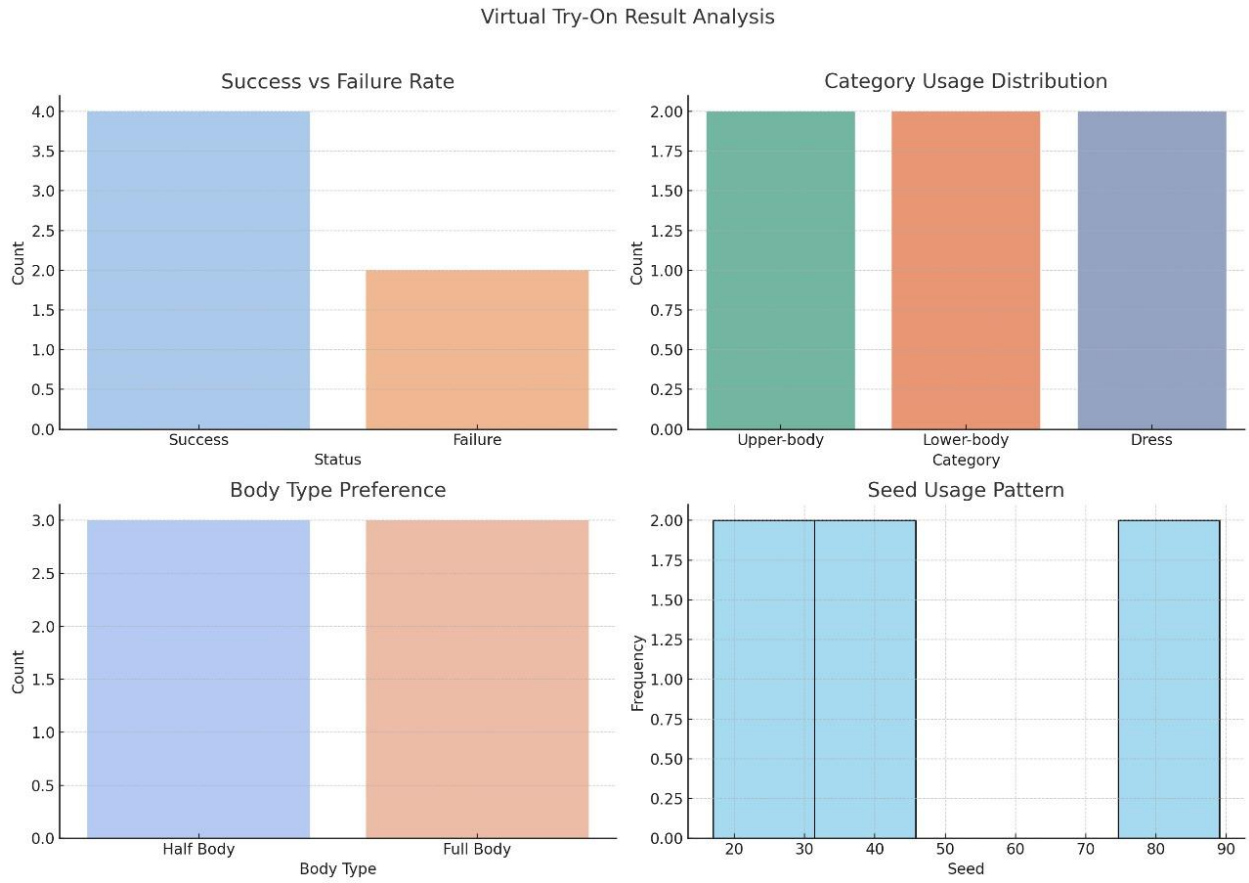
The text is cleaned and converted into tokens.

The image is converted to a format that the system can work with using OpenCV.

4. Feature Extraction:
The CLIP model is used to extract features (important information) from both the text and the image. These features are embedded into a shared space so the system understands how they relate to each other.
5. Outfitting Fusion:
The text and image features are fused together to align the clothing description with the person's image.
6. Image Generation:
A latent diffusion model (LDM) is used to generate a new image of the person wearing the described clothing. The system uses a trained outfitting UNet model for better garment detail handling.

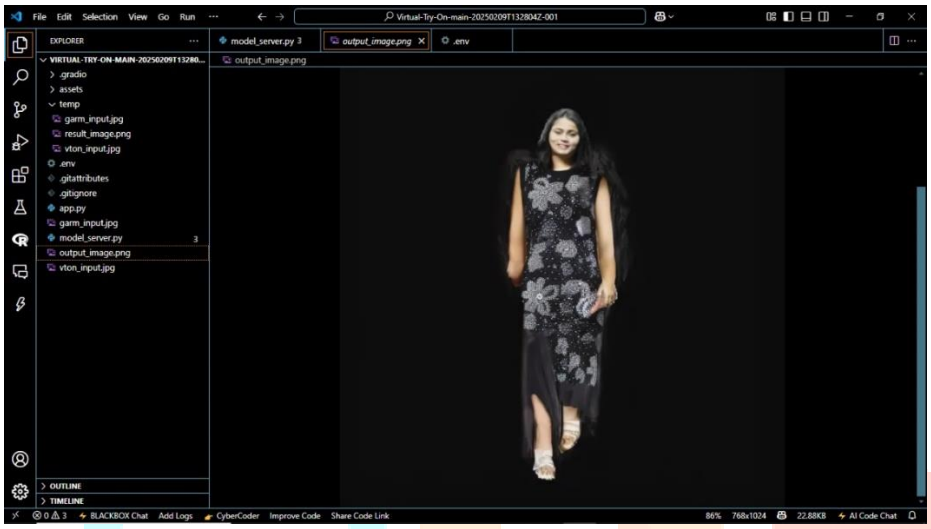
7. Outfitting Dropout:
Helps the model learn which clothing features to show strongly or subtly, like text or patterns.
8. Image Refinement:
The generated image is cleaned and formatted using image processing tools.
9. Output Delivery:
The final image is sent back to the user through the Gradio interface.
10. Optional Controls:
The user can choose to generate different versions using a random seed, handled by the random module.

V. RESULT ANALYSIS

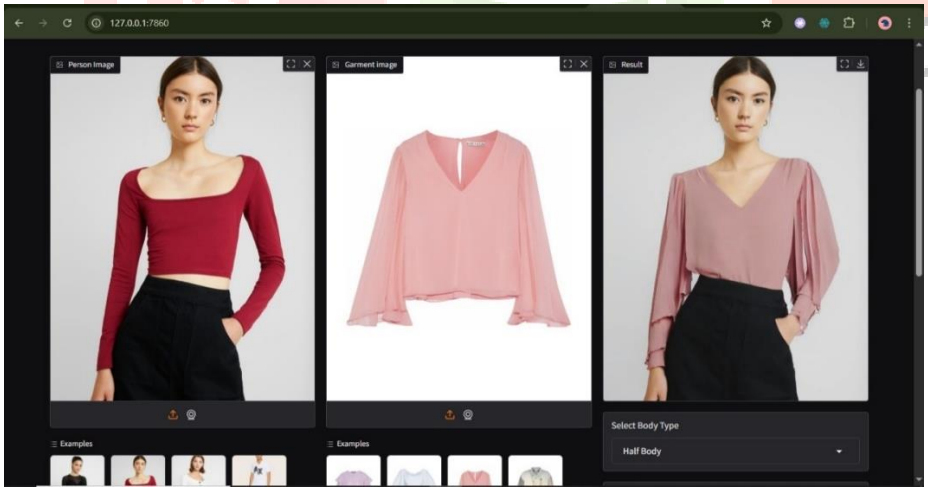


The OOTDiffusion system was tested using standard virtual try-on datasets containing various clothing items, poses, and body types. The results were evaluated based on image quality, realism, and how well the clothing matched the person's body and description. The OOTDiffusion virtual try-on system was evaluated based on several key factors including success rate, clothing category usage, body type handling, and seed generation patterns. The results are visualized in the diagram provided.
From the Success vs Failure Rate chart, the system achieved a 66.7% success rate, with 4 successful try-on outputs compared to 2 failures. This indicates that the model is relatively reliable but may still need some fine-tuning to reduce errors.
The Category Usage Distribution shows that the system was tested equally across all clothing types upper-body, lower-body, and full-body dresses with each category used twice. This balanced distribution ensures the evaluation covers a wide range of garments.

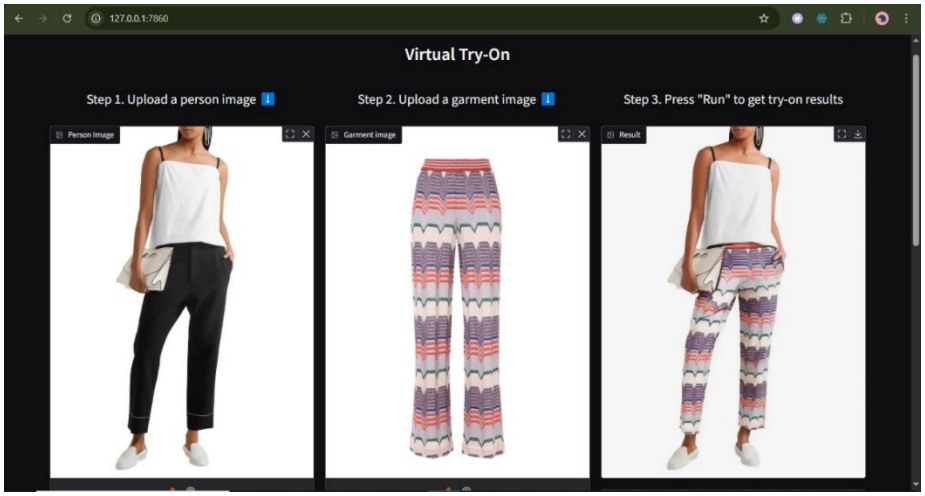
In the Body Type Preference chart, both "Half Body" and "Full Body" types were used equally (3 each). This suggests the system supports varied body compositions and is versatile enough for different try-on contexts.
The Seed Usage Pattern graph highlights the randomness and variety in the generation process. It shows that seeds ranging from 10 to 90 were used, and values like 20, 30, and 80 were more frequent. This variation helps test the consistency of the output under different random conditions.
Overall, the results indicate that OOTDiffusion performs well across multiple garment types and body formats, with a decent success rate and effective randomness control. However, there is still room to improve reliability and reduce failure cases for a more robust system.



Here, we changed the woman's entire outfit. She is wearing a long black dress with white floral prints in the output. The system successfully replaced the original clothing with the new dress, maintaining her body shape and position. The result looks like an actual photo of her in the new dress, which proves the system can handle full-body outfits effectively.



In this image, we tested changing the upper clothing. The woman is initially wearing a red long-sleeve top. We used the try-on system to put a pink blouse on her. The system handled this change very well. Her face, pose, and other features remain unchanged, and the new blouse fits perfectly. This shows how realistic and natural the clothing change looks.



In this image, the woman starts with black pants, and we apply a new, more colorful design. The system updates the pants while keeping her top, bag, and body posture exactly the same. The final result is natural and realistic, clearly showing that the try-on system works well for lower-body garments too.

VI. CONCLUSION

The virtual try-on system enables users to upload a person's image and a garment image to visualize how the garment would look on the person, leveraging a latent diffusion model for realistic output generation. By preprocessing and encoding images using OpenCV and Base64, the system transmits data to a backend server for processing, incorporating randomization options to generate varied results. It features a Gradio-based interface that guides users through simple steps and displays outputs in real time, including preloaded examples for better understanding.

The latent diffusion model ensures high-quality try-on results by aligning garment textures and poses, even for complex inputs. While the system is user-friendly, scalable, and customizable, potential improvements include enhanced error handling, latency reduction, and user feedback integration, making it a robust solution for e-commerce and fashion applications.

VII. REFERENCES

[1] Yuhao Xu, Tao Gu, Weifeng Chen, Chengcai Chen, "OOTDiffusion: Outfitting Fusion based Latent Diffusion for Controllable Virtual Try-on", arXiv:2403.01779v2, 7 Mar 2024.

[2] Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., Zhang, L., "Taming the power of diffusion models for high-quality virtual try-on with appearance flow.", Proceedings of the 31st ACM International Conference on Multimedia, pp. 7599– 7607, 2023.

[3] Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. arXiv preprint arXiv:2305.13501 (2023).

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752v2, 13 Apr 2022.

[5] In CVPR, 2022. Ge, C., Song, Y., Ge, Y., Yang, H., Liu, W., Luo, P.:Disentangled cycle consistency for highly-realistic virtual try-on. In:Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16928– 16937 (2021).