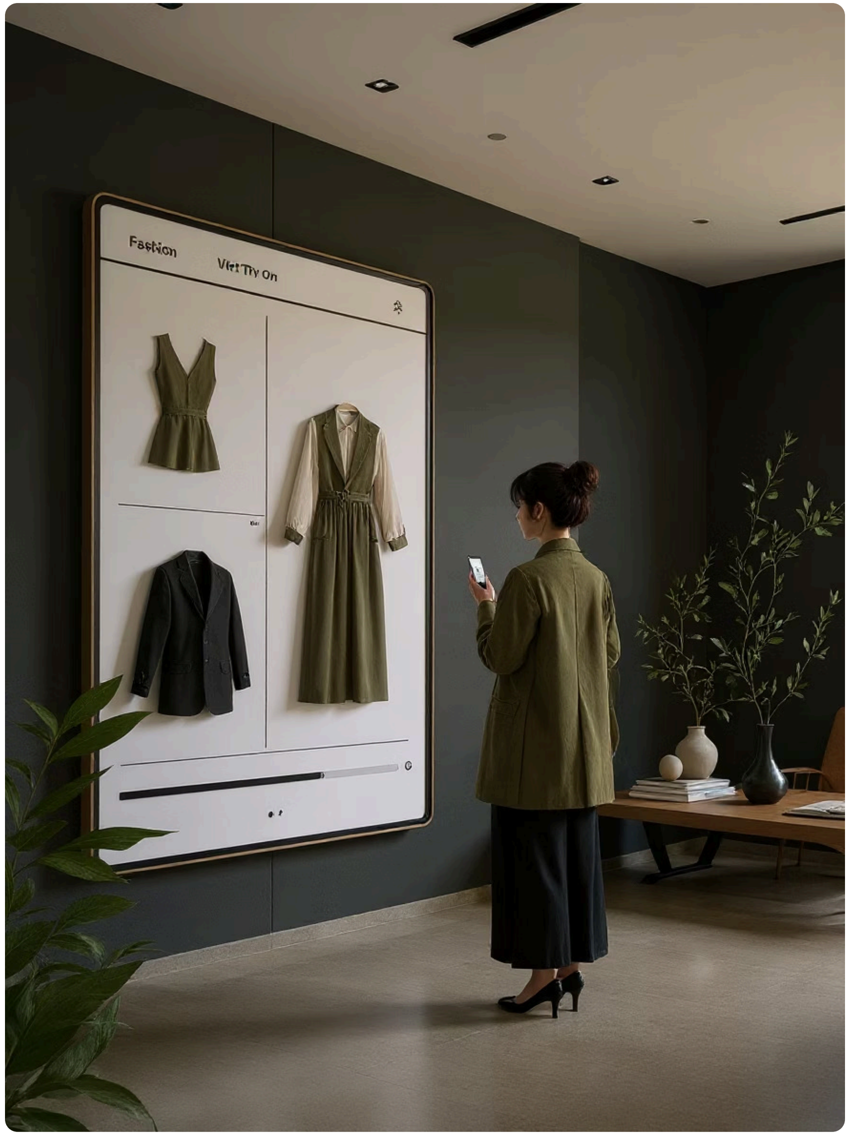


Virtual Try-On with Segment Anything Model and Diffusion

We will prototype a diffusion-based garment-transfer baseline model using mask-conditioned inpainting to replace apparel while preserving identity. After establishing fidelity and texture retention, we'll integrate SAM to generate high-precision garment masks, constraining edits to clothing regions only.

Group 14:

Chaitanya Kakade

Xinyi Shao

Riya Agrawal

Key Technical Hurdles in Realistic Virtual Try-On

Preserving Identity and Pose

The core objective is maintaining the person's identity, body shape, and pose while seamlessly integrating the virtual garment.

Natural Garment Deformation

The clothing product must be naturally deformed to fit the human body, accurately reflecting pose and body contours.

Detail Retention and Preservation

High-resolution details, such as fabric texture, logos, and intricate patterns, must be kept intact during the synthesis process.

Occlusion Handling

Body parts initially hidden by the original clothing in the input image must be properly rendered when the new garment is applied.

Limitations of Current Synthesis Approaches

Initial Warping & Fusion

Most prior research involves two stages: warping the clothing to fit the human body, followed by pixel-level fusion and refinement.

Segmentation Module Addition

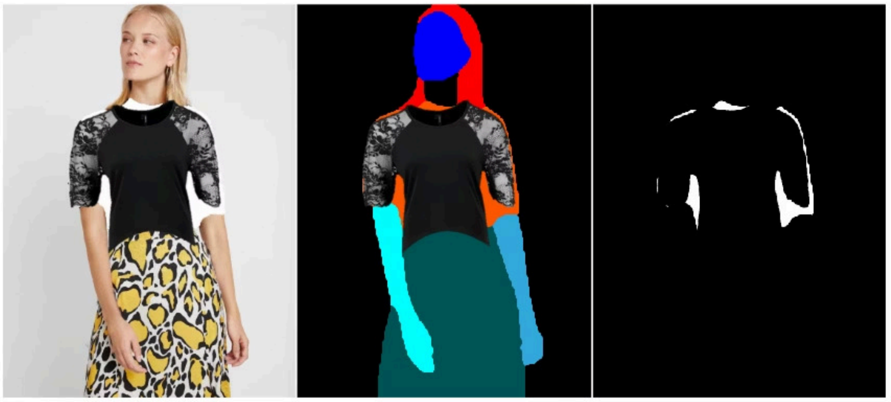
Recent methods (e.g., [9, 36, 35]) add a pre-processing module to generate segmentation maps, determining the person's final layout.

Thin-Plate Spline (TPS)

TPS is commonly used for deformation, though approaches like ClothFlow [9] also predict optical flow maps for better fit. However, this often fails to eliminate misalignment entirely due to regularisation, and is computationally costly.

The Resolution Artifact Problem

A significant drawback is the inability to scale to high resolution. As image size increases, artifacts and misalignments become visibly apparent, necessitating low-resolution output to mask these flaws.



(a) Warped Clothes on Reference Image

(b) Warped Clothes on Segmentation

(c) Misaligned Regions

Scaling up traditional methods reveals critical flaws, particularly misalignment between the warped clothing and the human body structure.

Better Approach: Enhanced Garment Agnostic Representation

New Clothing Agnostic Person Representation

A novel representation thoroughly eliminates all original clothing information by leveraging pose coordinates and a high-quality segmentation map.

Stable High-Resolution Synthesis

Instead of relying on a simple U-Net structure which leads to unstable training and unsatisfactory results, the improved method uses the new pose-based representation.

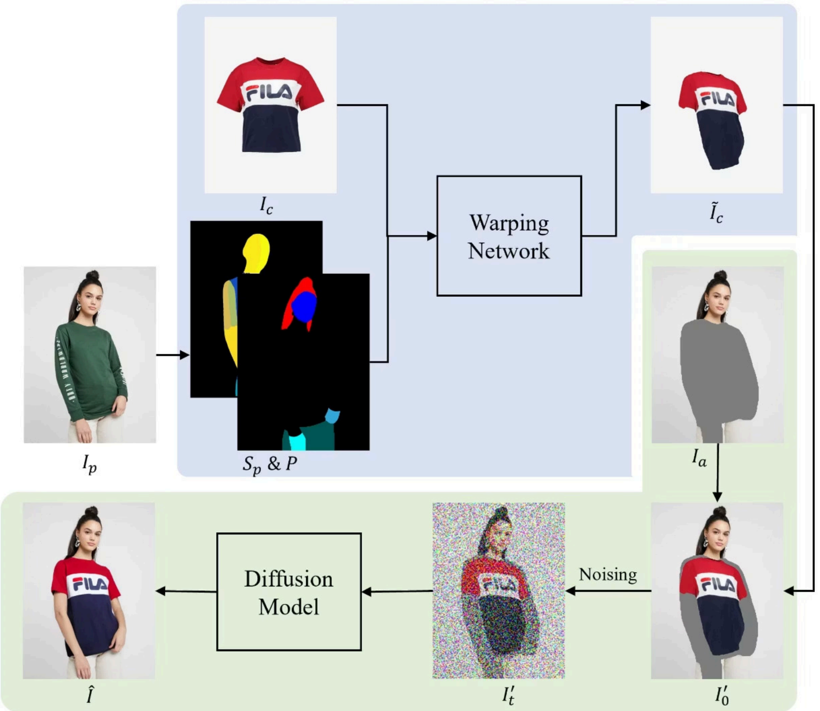
Refined Input Pipeline

The model is fed the detailed segmentation map and the target clothing item, pre-deformed to the human body, enabling superior high-resolution detail preservation.

This architectural shift provides a robust framework for high-fidelity virtual try-on, ensuring stable training and superior visual quality essential for commercial applications.

The improved solution:

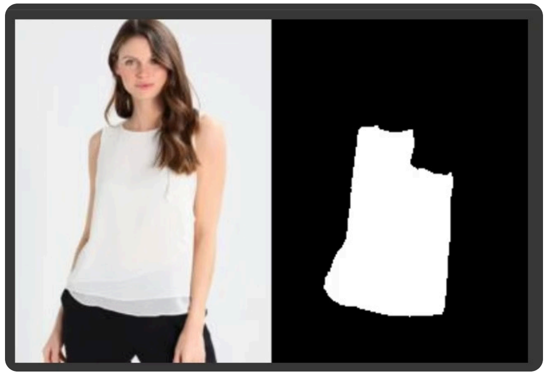
Using SAM and Diffusion



The overview of general method.

- Obtain the segmentation result , densepose and clothes-agnostic of the target person image through preprocessing.
- The clothes image is roughly aligned to the person by the warping network.
- We combine and ̃ to obtain ′o and add noise to get ′ as input to the diffusion model, and the final output ˆ produced by denoising ′ .

SAM Output



Stable Diffusion Output