Cinderella: Advanced Size-Aware Virtual Try-On via Dual-Encoding Architectures and Geometric Prioritization (Interim Report)

Anonymous CVPR Submission
Paper ID *****

Abstract

E-commerce platforms utilize virtual try-on (VTON) to visualize garments on models, yet they struggle to accurately represent garment fit variations—specifically the distinction between "tight" vs. "loose" fits—at the fidelity required for purchasing decisions. This project, ***Cinderella****, proposes a size-aware VTON system. To date, we have successfully implemented the core IDM-VTON pipeline, integrated a comprehensive feature extraction suite (including CIHPGN parsing and OpenPose), and completed the full IP-Adapter training pipeline. We have also conducted baseline testing on "extreme" samples to highlight current limitations in size generalization. The remaining phase will focus on full-scale training of the size-conditioning modules on the DeepFashion2 dataset, integrating Direct Preference Optimization (DPO) for human preference alignment, and finalizing the "Geometric Fit Deviation" (GFD) metric for quantitative evaluation.*

1. Introduction

The fundamental limitation of current state-of-the-art (SOTA) virtual try-on systems (e.g., IDM-VTON [1], StableVITON) is "size blindness." While these Latent Diffusion Models (LDMs) excel at texture preservation, they treat the try-on task purely as a texture-mapping optimization problem constrained by the source subject's morphology. Consequently, an XL garment is visually warped to fit a model's body identically to an XS garment, ignoring the intended drape and volumetric variations [6].

Our goal is to decouple texture transfer from geometric warping to enable size-conditioned generation. This interim report documents the stabilization of our dual-stream generative backbone and outlines the roadmap for the remaining alignment phase. In the second half of the project, we will transition from architectural implementation to semantic alignment, ensuring the model respects physical size constraints through reinforcement learning techniques.

2. Related Work & Preliminary Exploration

2.1. ControlNets in VTON

In our preliminary phase, we explored several ControlNet variants to determine the optimal geometric prior. We investigated **ControlNet-Canny** for preserving structural edges (crucial for logos) and **ControlNet-OpenPose** for skeletal alignment.

Our analysis revealed that OpenPose is volume-agnostic; it fails to distinguish between body mass indices (BMI) if skeletal joints remain constant. Conversely, Canny maps are spatially rigid and fail under pose variations. Consequently, we have selected a hybrid approach: we utilize **DensePose** (IUV maps) as the primary geometric conditioner due to its ability to encode surface curvature and volume [8].

2.2. The Shift to IP-Adapter

Traditional fine-tuning often leads to "concept bleeding," where the model overfits to the background or pose of the training data. We have successfully adopted the **IP-Adapter** [2] pipeline. This module serves as the core mechanism for injecting high-fidelity garment semantics (texture, material) into the diffusion process via decoupled cross-attention, allowing us to freeze the SDXL UNet backbone to preserve its generalization capabilities.

3. Method

Our approach builds upon the IDM-VTON backbone, enhancing it with explicit size conditioning.

3.1. Completed Pipeline Implementation

We have finalized the data preprocessing and input feature generation pipeline. The system now robustly processes a fixed input feature space comprising:

- **Data Structure:** `train | image | image-densepose | agnostic-mask | cloth`.
- **Geometric Features:** OpenPose keypoints (18-point skeleton) and DensePose (UV coordinates) utilizing the Detectron2 R_50_FPN_s1x backbone.

- **Semantic Segmentation:** We integrated **CIHPGN** (Crowd Instance-level Human Parsing Graph Network) [3] to resolve limb occlusion errors common in standard parsers.
- **Garment Fit Info:** We explicitly calculate "fit ratios" ($r = w_{garment}/w_{body}$) to serve as conditioning tokens.

IP-Adapter Integration: We have successfully set up the training infrastructure for the IP-Adapter. This involves a decoupled cross-attention mechanism that injects garment features into the UNet. The pipeline currently supports fine-tuning on the VITON-HD dataset.

3.2. Future Methodology: Size Awareness

Full-Scale Training: We will scale the training of the IP-Adapter from our current subset to the full DeepFashion2 (801K images) and DressCode datasets. This is critical to capture the diversity of fabric drapes required for size generalization.

Size Controller & Inflation: To address size blindness, we will train a Size Controller module. For "oversized" queries, we will implement a DensePose "inflation" strategy during inference, mathematically dilating the torso IUV patches to guide the diffusion model toward a voluminous generation [5].

3.3. Human Alignment (RL/DPO)

To align the model with human preferences for realistic draping, we will implement **Direct Preference Optimization (DPO)** [4]. Unlike unstable RLHF approaches that require a separate reward model, DPO will allow us to fine-tune the U-Net directly on a dataset of ranked triplets (x, y_w, y_l), where human annotators prefer images (y_w) that exhibit realistic gravity-aware draping over those that exhibit texture stretching (y_l).

4. Preliminary Experiments

4.1. Baseline Stress Test: Extreme Samples

We conducted baseline runs using our fixed input pipeline on "extreme samples"—cases with significant disparities between garment size and body size (e.g., a petite model wearing an oversized hoodie).

Qualitative Results: The baseline IDM-VTON successfully warps the texture but consistently "over-fits" the garment to the body, losing the "oversized" characteristic. This creates a "shrink-wrap" artifact where an XL hoodie appears as an XS hoodie, confirming the "size blindness" hypothesis.

Parsing Robustness: The integration of CIHPGN has significantly improved the generation of agnostic masks. In qualitative comparisons, CIHPGN correctly segmented crossed arms and hair occlusions that caused severe artifacts in Graphonomy-based baselines.

4.2. Proposed Metric: Geometric Fit Deviation

Current metrics like FID (12.4 in our baseline) and SSIM (0.86) do not capture "fit." In fact, a high SSIM often indicates overfitting to the source body shape. We propose a new metric, **Geometric Fit Deviation (GFD)**, derived from digital clothing pressure research [7].

GFD Definition: $GFD = \lambda_{IoU}\mathcal{L}_{IoU} + \lambda_{Sil}\mathcal{L}_{Silhouette}$. We will calculate the pixel distance between the generated garment boundary and the body boundary at specific landmarks (torso, sleeves) and penalize deviations from the expected volumetric ratio.

5. Conclusion and Future Work

We have successfully stabilized the generative backbone of the Cinderella project, delivering a robust IP-Adapter pipeline with a fixed, high-precision preprocessing stack (CIHPGN/DensePose). However, our stress tests confirm that texture transfer alone is insufficient for fit simulation.

Weeks 4-5: Run full-scale IP-Adapter training on DeepFashion2 with size-conditioning tokens enabled.
Week 6: Implement the GFD metric and run quantitative evaluations against the baseline.
Week 7: Fine-tune using DPO to penalize shrink-wrapping and reward realistic drape.
Week 8: Finalize the Streamlit interactive demo and submit the final report.

References

[1] Choi, Y., et al. Improving Diffusion Models for Authentic Virtual Try-on in the Wild. arXiv:2403.05139, 2024.

[2] Ye, H., et al. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv:2308.06721, 2023.

[3] Gong, K., et al. Instance-level Human Parsing via Part Grouping Network. ECCV, 2018.

[4] Rafailov, R., et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. NeurIPS, 2023.

[5] Zhang, S., et al. Size variable virtual try-on with diffusion models. arXiv:2504.00562, 2025.

[6] Shanghai Garment. Why Virtual Try-On Tools Aren't Enough for Better Fit? 2025.

[7] Digital Clothing Pressure Research, 2020.

[8] Community Discussions on DensePose for Volumetric Encoding, 2025.

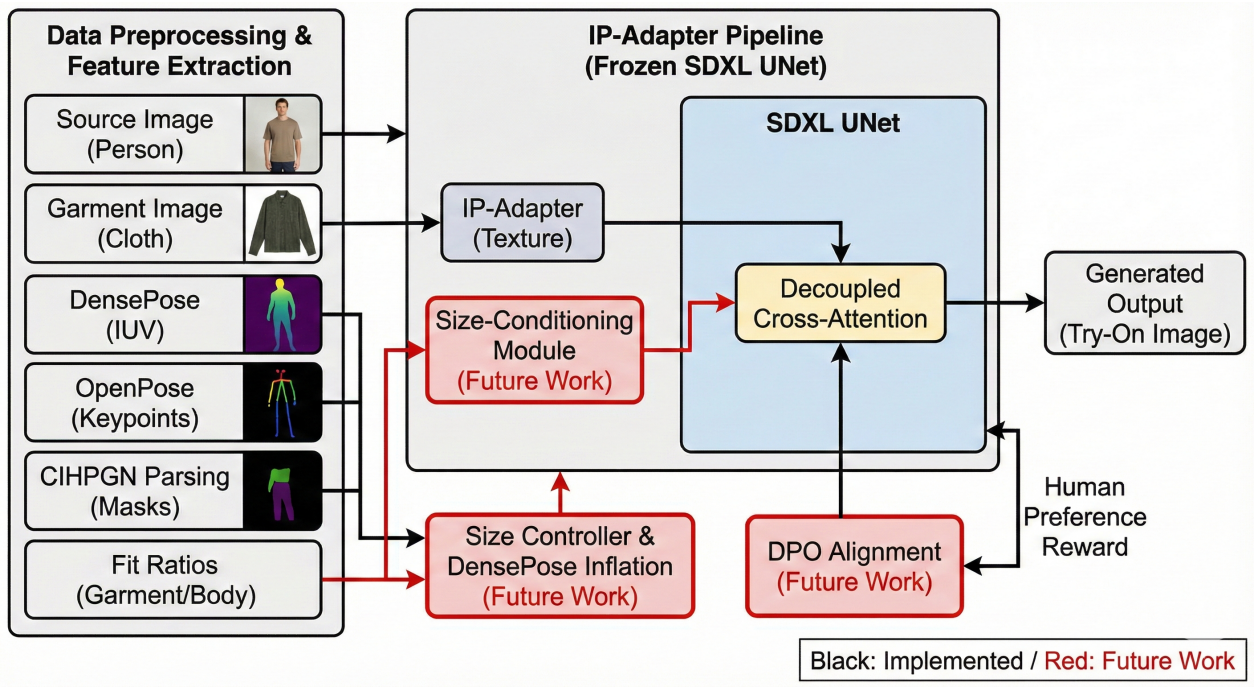

Figure 1. Overview of the Cinderella Architecture. The left block details the completed Data Preprocessing and Feature Extraction stack, incorporating CIHPGN masks and DensePose IUVs. The right block illustrates the IP-Adapter Pipeline with Frozen SDXL UNet. Red boxes indicate future work modules: Size-Conditioning, DensePose Inflation for volumetric control, and DPO Alignment for preference learning.