Cinderella: Advanced Virtual Try-on

Anonymous CVPR submission

Paper ID *****

Abstract

E-commerce platforms can generate virtual try-on images but cannot accurately represent garment fit variations (tight vs. loose) at the same detail fidelity required for purchase decisions. Our approach combines IDM-VTON's dual-encoding architecture (IP-Adapter for semantics, GarmentNet for low-level details) with a novel size-conditioning module that learns garment-to-body dimensional ratios, producing per-image try-on results where an XL garment appears appropriately loose on a small frame. Fine-tuning a fashion-specific IP-Adapter and integrating spatial size guidance maps addresses the core challenge: preserving intricate garment patterns while respecting physical size constraints. The deliverable includes a 2D try-on pipeline with optional multi-view 3D rendering, comprehensive evaluation metrics, and an interactive demo for real-world validation. We propose a comprehensive virtual try-on addressing three aspects: detail preservation, size aware generation and 3D controllable rendering to address the critical limitations

1. Introduction

When shoppers evaluate garments online, they face a critical question that existing virtual try-on (VTON) systems fail to answer: *will this size actually fit me?* Current VTON methods excel at warping garments to body contours but ignore the fundamental relationship between garment dimensions and body proportions—an XL shirt renders identically fitted whether the model is petite or plus-size. This limitation stems from treating try-on purely as a texture transfer problem rather than a size-aware transformation task. Our goal is a learning-based system that generates authentic virtual try-on images with explicit size control, where garment dimensions determine visual fit characteristics (draping, tightness, coverage area). Each output explains how the size was determined through interpretable size maps and attention visualizations, enabling both automated sizing recommendations and human verification.

2. Limits of Current Practice

Today's VTON systems face three persistent challenges. **First, size blindness:** models like HR-VITON, VITON-HD, and even recent diffusion methods (LaDI-VTON, DCI-VTON, StableVITON) warp garments to fit target poses without considering whether a garment is XS or XXL—the same hoodie appears perfectly fitted on all body types. **Second, detail loss:** while GAN-based methods struggle with pattern preservation, diffusion models often blur logos, text, and intricate designs when conditioning is insufficient. **Third, lack of controllability:** users cannot specify "show me this in a size larger" without re-running with different garment images. Prior work addresses detail (IDM-VTON's GarmentNet) or size (COTTON's landmark-based warping) separately, but no system combines both with diffusion model quality. This gap motivates our integrated design.

3. Approach and Key Insight

We treat size-aware try-on as a multi-modal conditioning problem: garment appearance (preserved via dual encoding), body geometry (captured via DensePose and pose keypoints), and size relationship (learned from garment-body dimension ratios). By injecting size information into attention mechanisms rather than relying solely on implicit warping, the model learns when to generate tight-fitting versus loose-draping results. Fine-tuning IP-Adapter on fashion-specific semantics (materials, styles, necklines) further ensures that "cotton t-shirt" differs meaningfully from "silk blouse" in rendered texture.

3.1. Data and Size Annotation Policy

Base Datasets. VITON-HD (11,647 pairs, 1024×768) for primary training; DressCode (multi-category) for cross-category generalization testing; DeepFashion2 (801K images, 13 categories, 491K clothing items with bounding boxes, dense landmarks, and commercial consumer-to-shop pairs) for fashion-specific semantic understanding and landmark detection pre-training; custom In-the-Wild set (1,500+ pairs) with size annotations for size-aware evaluation. DeepFashion2's rich annotations including clothing landmarks, scale, occlusion, zoom-in, viewpoint, and bounding boxes provide ideal training data for our 10-point garment landmark predictor and material/style classification components.

Size Labeling. We extract garment dimensions (shoulder width, torso/sleeve length) via 10-point CNN landmark predictor pre-trained on DeepFashion2 and fine-tuned on 500 VITON-HD samples, and body dimensions from pose keypoints (OpenPose/MediaPipe). Size ratio $r = (\text{garment/body width}, \text{garment/body length})$ maps to discrete labels: tight ($r < 0.9$), fitted ($0.9 \leq r < 1.1$), loose ($1.1 \leq r < 1.3$), oversized ($r \geq 1.3$). 500 images manually verified in Label Studio; augmentation scales garments 0.7–$1.5\times$ to balance distribution.

3.2. Preprocessing

Person Image Processing. (1) Human parsing: SCHP/Graphonomy produces 20-class segmentation. (2) DensePose estimation: Detectron2 extracts UV body surface maps. (3) Pose keypoints: OpenPose provides an 18-point skeleton. (4) Agnostic mask: Remove original garment while preserving arms (to retain skin tone, tattoos, width). (5) Encoding: All inputs (masked person, DensePose, mask) passed through VAE encoder to latent space ($128 \times 96 \times 4$).

Garment Image Processing. (1) Segmentation: SAM/U^2-Net isolates garment from background. (2) Landmark detection: 10-point predictor (neck, shoulders, elbows, wrists, hips) trained on a curated set. (3) Caption generation: Fashion-specific CLIP tagger extracts: sleeve type, neckline, material, features → Template: "A photo of [attributes]". (4) Encoding: Garment latent ($128 \times 96 \times 4$) + CLIP image features (257×1280).

3.3. Model Architecture and Training

IDM-VTON Core (Base). Our base architecture consists of three components: *TryonNet*, an SDXL Inpainting UNet with 13-channel input (noised latent + mask + masked person + DensePose); *GarmentNet*, a frozen SDXL UNet encoder extracting multi-scale garment features (low-level details: patterns, logos, textures); and *IP-Adapter*, combining frozen CLIP ViT-H/14 with trainable projection layers for high-level semantics (style, color, material).

Size Module (Novel). We introduce two novel components for size awareness: *Size Encoder*, an MLP mapping size ratio $[\text{width\_ratio}, \text{length\_ratio}, \text{sleeve\_ratio}] \rightarrow$ 768-dim embedding; and *Size Controller*, a CNN generating spatial size maps (per-pixel tight/loose guidance) from fused person + garment + size features.

Size-Aware Attention. We modify self/cross-attention layers to incorporate size information. *Self-attention* modulates token importance based on size maps (loose regions attend more to garment structure), while *cross-attention* injects size embeddings alongside text/image features.

3.4. Training Protocol

Stage 1: Base IDM-VTON. We use pretrained TryonNet and fine-tune IP-Adapter projection on VITON-HD while freezing GarmentNet to preserve SDXL prior knowledge.

Stage 2: IP-Adapter Fashion Fine-tuning (30 epochs). We freeze TryonNet/GarmentNet and fine-tune IP-Adapter on 5,000 fashion images, adding custom attention processors with garment-specific gating.

Stage 3: Size Module Training (50 epochs). We freeze the base IDM-VTON and train the size encoder/controller using an augmented dataset with synthetic size variations. The loss function combines reconstruction, size consistency, and spatial alignment: $\mathcal{L}_{\text{rec}} + 0.5 \cdot \mathcal{L}_{\text{size\_consistency}} + 0.3 \cdot \mathcal{L}_{\text{spatial}}$.

Stage 4: Joint Fine-tuning (30 epochs). We unfreeze the TryonNet decoder, IP-Adapter, and size modules for end-to-end optimization with multi-objective loss: $0.3 \cdot \mathcal{L}_{\text{idm}} + 0.25 \cdot \mathcal{L}_{\text{ip}} + 0.25 \cdot \mathcal{L}_{\text{size}} + 0.15 \cdot \mathcal{L}_{\text{detail}} + 0.05 \cdot \mathcal{L}_{\text{human}}$.

Stage 5: GS-VTON 3D (20 epochs, optional). We optionally use 2D outputs as supervision for 3D Gaussian splatting scenes, optimizing Gaussian parameters (position, color, opacity) to match multi-view try-on results.

4. Implementation Details

Hardware and Training Time. Our system requires $4\times$ A100 (80GB) or H100 GPUs with total training time of approximately 70 hours: Base (25h), IP-Adapter (15h), Size (20h), and Joint fine-tuning (10h).

Development Timeline. *Week 1:* Data curation (size annotations, landmark training), infrastructure setup, and baseline IDM-VTON implementation. *Week 2:* IP-Adapter fashion fine-tuning with custom processors and material classification. In parallel, size module development (encoder, controller, attention modifications) and integration testing. *Weeks 3–4:* Joint fine-tuning and ablation studies (w/o size, w/o IP-Adapter tuning, w/o GarmentNet). *Week 5:* Optional GS-VTON 3D extension and multi-view consistency. *Week 6:* Evaluation, user studies, and demo development (Streamlit interface).

5. Evaluation and Success Criteria

5.1. Mid-Term Evaluation

We evaluate our model on held-out validation scenes (DressCode + 200 In-the-Wild images) using three criteria:

Detail Preservation. We require LPIPS < 0.12 compared to the IDM-VTON baseline (≈ 0.102) to ensure our size-aware modifications do not degrade visual quality.

Size Awareness Baseline. We compare against two baselines: (a) IDM-VTON without size conditioning, and (b) a

Table 1. Quantitative evaluation metrics and targets. Our model aims to match or exceed the baseline IDM-VTON while introducing size awareness capabilities.

Metric	Target	Baseline (IDM-VTON)
LPIPS (Detail)	< 0.10	0.102
SSIM (Structure)	> 0.90	0.870
FID (Realism)	< 6.0	6.29
Size Accuracy	$> 85\%$	NA
CLIP-I (Similarity)	> 0.90	0.883
DISTS	> 0.85	—
LPIPS-Clothing	< 0.08	—
LPIPS-Person	< 0.12	—
mIoU	> 0.75	—
FTS	> 0.80	—
Size Accuracy (Ratio)	$> 85\%$	—

size-only heuristic that scales garments by dimension ratio and performs simple paste operations.

Qualitative Attention Verification. Using Grad-CAM visualization, we verify that attention mechanisms focus on garment boundaries rather than background for 90%+ of samples, as rated by two annotators on 100 patches.

5.2. Final Evaluation

Quantitative Metrics. We evaluate on a test set of 500 images using the following metrics:

References

[1] Yukang Cao, Masoud Hadi, Liang Pan, and Ziwei Liu. Controllable 3d virtual try-on with gaussian splatting, 2024. arXiv preprint arXiv:2410.05259.

[2] Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, and Wen-Huang Cheng. Size does matter: Size-aware virtual try-on via clothing-oriented transformation. In *ICCV*, 2023.

[3] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *ECCV*, 2024.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

[5] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. arXiv preprint arXiv:2307.01952.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[7] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. arXiv preprint arXiv:2308.06721.

[8] Shufang Zhang, Hang Qian, Minxue Ni, Yaxuan Li amd Wenxin Ding, and Jun Liu. Size variable virtual try-on with diffusion models, 2025. arXiv preprint arXiv:2504.00562.