WEARANYTHING: DRESS 3D HUMANS IN SECONDS VIA DECOUPLED GAUSSIAN GARMENT TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

We present WearAnything, a feed-forward framework for direct 3D garment transfer in Gaussian human representations, enabling efficient and controllable virtual try-on without multi-stage reconstruction or subject-specific optimization. Existing 3D virtual try-on methods are computationally expensive and often struggle with cross-category transfer involving substantial changes in garment coverage. WearAnything performs garment transfer directly in the canonical token representation of a pretrained 3D human model. The framework learns separate identity and garment latent spaces from dense canonical features, while a lightweight routing module adapts the editable region to the target garment support. Together, these components enable the recomposition of source identities and target garments across clothing categories, including transfers with substantial changes in garment coverage. To maintain 3D structure during garment editing, we introduce a 3D distillation strategy that transfers canonical feature and Gaussian geometry supervision from the frozen reconstruction model to the editable representation. The learned garment space further supports continuous outfit interpolation. Experiments demonstrate state-of-the-art 3D garment transfer quality and fast end-to-end inference within a few seconds.

1 INTRODUCTION

3D virtual try-on aims to transfer a desired garment onto a target human while preserving identity, body shape, pose, and view-consistent appearance. Existing 2D methods produce high-quality edits but lack an explicit 3D representation for novel-view rendering and avatar applications. Recent 3D methods address this gap through subject-specific reconstruction and editing pipelines. LayGA Lin et al. (2024), for example, reconstructs the human body and clothing as separate Gaussian layers from multi-view videos before transferring a fully modeled garment layer to another reconstructed avatar. GS-VTON Cao et al. (2026) reconstructs a subject-specific 3D Gaussian representation, applies a personalized 2D diffusion model to inpaint garment regions in rendered views, and then updates the 3D representation using the edited observations. However, these pipelines require several stages of reconstruction, editing, and optimization for each subject or garment, resulting in high computational cost. Moreover, diffusion inpainting modifies a garment region defined on the existing avatar rather than directly restructuring its 3D garment support. This editing approach is most effective when transferring garments within the same category and with similar spatial coverage. Cross-category transfer, such as replacing a short skirt with long trousers, requires both removing source-garment geometry and synthesizing target geometry in previously uncovered regions, making it difficult to achieve reliably with local appearance inpainting alone.

To address these challenges, we present **WearAnything**, a feed-forward framework for direct garment transfer in 3D Gaussian human representations. WearAnything operates directly on the canonical token representation of a pretrained human reconstructor, avoiding per-subject reconstruction and iterative editing while substantially reducing inference time. To enable controllable garment transfer in this space, we introduce a lightweight identity and garment decoupling design with learned garment routing. Cross-category transfer, however, requires more than separating clothing from identity because the model must also represent garment geometry and change its spatial coverage. We therefore learn a dedicated garment latent space that captures garment appearance, geometry, and coverage, and use routing-aware composition to combine the target garment with the source identity. This design adds or removes garment features according to the target support while preserving source

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

body shape and non-clothing appearance. We further introduce a Decoupling Training Strategy with 3D distillation to strengthen identity and garment separation while preserving the 3D geometric consistency of the transferred avatar. Extensive experiments demonstrate geometrically consistent cross-category transfer under large changes in garment coverage while preserving source identity. Beyond direct transfer, the continuous garment latent space enables smooth interpolation between outfits, a capability not supported by existing 3D virtual try-on pipelines.

Our contributions are summarized as follows:

- We introduce **WearAnything**, a feed-forward 3D virtual try-on framework that performs garment transfer directly in a Gaussian human representation within a single forward pass.
- We propose a lightweight identity and garment decoupling module with routing-guided composition, which separates source identity and body information from target garment appearance, geometry, and coverage in dense point-wise token space, enabling localized and identity-preserving transfer.
- We introduce a decoupling training strategy with 3D feature and geometry distillation, which strengthens the separation between identity and garment representations while preserving structural validity and view-consistent geometry during transfer.
- We demonstrate that WearAnything achieves fast and identity-preserving 3D virtual try-on while learning a continuous garment latent space that supports smooth outfit interpolation and controllable garment editing.

2 Related Work

Early 3D virtual try-on methods largely follow an image-to-3D reconstruction paradigm, including template-based surface transfer, monocular depth reconstruction, implicit surface modeling, and diffusion-guided 3D optimization Lei et al. (2020); Zhao et al. (2021); Du et al. (2022); Hu et al. (2023); Li et al. (2024); Zhong et al. (2025a); Xie et al. (2024); Chen et al. (2024); Cao et al. (2026). A related line of work develops garment aware avatar representations by explicitly separating the body and clothing layers for reconstruction, animation, and garment transfer Lin et al. (2024); Chen et al. (2026); Xu & Dingliana (2026); Jin et al. (2026); these methods primarily focus on constructing editable layered avatars rather than direct feed-forward virtual try-on from a source person and reference garment.

More recently, AnyView VTON Zhang et al. (2025) formulates 3D virtual try-on as view-conditioned multi-view diffusion editing followed by Gaussian reconstruction, improving appearance consistency across viewing directions. G3D-VTON Xie et al. (2025) employs a conditional 3D-aware generative model with semantic articulated fields to obtain view-controllable try-on representations, followed by image-space refinement for appearance fidelity. VTON 360 He et al. (2025) reformulates 3D virtual try-on as multi-view 2D try-on using pseudo-3D pose guidance, multi-view spatial attention, and camera-aware garment features before lifting the generated views into 3D. IPVTON Zhong et al. (2025b) optimizes a hybrid 3D human representation using image-conditioned score distillation together with mask-guided garment conditioning and silhouette-based geometric constraints. Despite substantial progress in garment fidelity and multi-view consistency, most recent approaches still construct the final 3D result through multi-view generation, diffusion-guided optimization, or subsequent 3D reconstruction. In contrast, WearAnything performs garment transfer directly in the canonical token representation of a pretrained Gaussian human reconstructor, enabling a single feed-forward transfer without per-subject optimization or multi-view lifting.

3 Method

Given a source human image I_A and a reference outfit image I_B, our goal is to reconstruct a 3D Gaussian avatar that preserves the identity, body shape, and pose of I_A while wearing the garment from I_B. WearAnything follows a feed-forward framework that performs garment transfer directly in the canonical token space of a frozen Gaussian human reconstructor. Instead of editing rendered images and reconstructing the result, we predict a point-wise garment routing mask and decouple identity and garment information within each canonical token representation. We then combine the

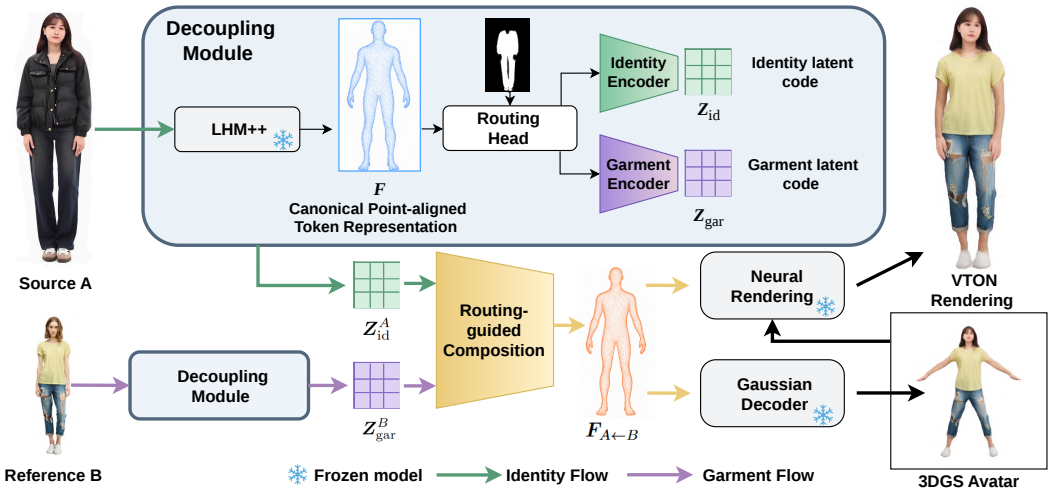


Figure 1: **Pipeline of WearAnything**. WearAnything is a feed-forward framework for direct garment transfer in the canonical token space of a pretrained 3D human model. Given a full-body source image and reference image of a person wearing the target outfit, the frozen reconstructor maps both inputs to point-aligned dense canonical features. Separate encoders decouple these features into identity and garment latent spaces, and a composition module combines the source identity with the reference garment. A learned garment routing mask defines the target garment support in canonical space, allowing the composed representation to add or remove garment regions while preserving source body shape and non-clothing appearance. The composed tokens are then decoded into an explicit 3D Gaussian avatar and rendered views.

source identity with the reference garment through routing-guided feature composition and decode the composed field into a 3D Gaussian avatar. The shared canonical correspondence enables localized garment transfer, while 3D distillation from the frozen reconstructor preserves the geometric structure of the pretrained representation.

3.1 Gaussian Human Reconstructor

To enable feed-forward 3D garment transfer, we use LHM++ Qiu et al. (2025b) as our Gaussian human reconstructor. It is a feed-forward human reconstruction model that maps one or multiple input images $\{I_k\}_{k=1}^{K}$ through three representation spaces. First, its token extractor Φ produces a dense canonical point-aligned token representation.

$$\boldsymbol{F} = \Phi(\{I_k\}_{k=1}^{K}) \in \mathbb{R}^{N\times D}, \tag{1}$$

where N is the number of canonical human points and D is the token dimension. Each token $\boldsymbol{F}_n$ is associated with the n-th canonical point and encodes its local appearance and geometric information. Since different inputs share the same ordered canonical point set, tokens with the same index are in point-wise correspondence across images. From the shared canonical token representation $\boldsymbol{F}$, LHM++ first decodes a canonical 3D Gaussian representation. The Gaussian decoder $\mathcal{G}$ directly maps $\boldsymbol{F}$ to a set of 3D Gaussians $\mathcal{S}$,

$$\mathcal{S} = \mathcal{G}(\boldsymbol{F}), \tag{2}$$

where each Gaussian contains its position, scale, rotation, opacity, and appearance attributes. The decoded Gaussians then provide the geometry and visibility required for rendering. Gaussian rasterization can directly produce an RGB image from their appearance attributes, while also splatting the 128 dimensional token features into the image plane. The neural feature renderer $\mathcal{R}$ consumes this rasterized feature map and predicts the final RGB image and foreground mask,

$$(\hat{I}, \hat{\boldsymbol{M}}_{\text{fg}}) = \mathcal{R}(\boldsymbol{F}; \mathcal{S}, \mathcal{P}), \tag{3}$$

where $\mathcal{P}$ is the pose estimated from I. Internally, $\mathcal{R}$ poses the canonical tokens according to $\mathcal{P}$, projects and rasterizes their features using the Gaussian positions, scales, rotations, opacities, and visibility, and converts the resulting feature map into $\hat{I}$ and $\hat{\boldsymbol{M}}_{\text{fg}}$ using a Dense Prediction Transformer Ranftl et al. (2021).

3.2 DECOUPLED GARMENT TRANSFER

The canonical token representation $\boldsymbol{F}$ connects both the Gaussian representation $\mathcal{S}$ and the rendered image $\hat{I}$. This shared representation provides a natural interface for 3D garment transfer. However, LHM++ learns these tokens for reconstruction, leaving identity, body shape, garment geometry, and garment appearance entangled. Direct token replacement can therefore introduce the reference identity and alter non-clothing regions. To address this problem, we decouple identity and garment information and recombine them through routing-guided feature composition.

3.2.1 GARMENT ROUTING

Garment transfer requires identifying which canonical tokens describe clothing and which should retain source identity. This division cannot be fixed across inputs because different garments may occupy different body regions, such as shirts with varying sleeve lengths. We therefore condition the routing mask on an aligned image-space garment mask. For each input, the mask is projected onto the canonical tokens through the frozen Gaussian geometry, producing a garment prior $\boldsymbol{p}_X$ and a projection-validity indicator $\boldsymbol{v}_X$. We denote the frozen canonical representation by $\boldsymbol{F}_X = \Phi(I_X)$, where $X \in \{A, B\}$, and the canonical coordinates by $\boldsymbol{Q}$. A shared routing head R predicts base logits, while a lightweight graph refinement head H incorporates the token features, coordinates, and projected mask:

$$\boldsymbol{m}_X = \sigma(R(\boldsymbol{F}_X) + H(\boldsymbol{F}_X, \boldsymbol{Q}, \boldsymbol{p}_X, \boldsymbol{v}_X) + \mathrm{logit}(\boldsymbol{p}_X)) \in (0,1)^N. \tag{4}$$

Here, $\boldsymbol{m}_{X,n}$ is the garment probability of canonical point n, while invalid projections fall back to the learned routing prediction. Applying the same mask-conditioned routing to both inputs yields $\boldsymbol{m}_A$ and $\boldsymbol{m}_B$. During transfer, $\boldsymbol{m}_B$ determines the support of the inserted garment, allowing its coverage and silhouette to follow the reference outfit.

3.2.2 IDENTITY AND GARMENT DECOUPLING WITH FEATURE COMPOSITION

Directly transferring reference tokens would also carry reference-specific identity information, while overly compressing all clothing content would discard garment texture and geometry. Effective transfer therefore requires two representations with different roles. The identity representation should remain compact and stable across outfits, whereas the garment representation must retain sufficient capacity for detailed appearance, geometry, and coverage. This motivates routing-conditioned encoders with asymmetric latent bottlenecks. Given $\boldsymbol{F}_X$ and its routing mask $\boldsymbol{m}_X$, two encoders extract the identity and garment representations

$$\boldsymbol{Z}^X_{\mathrm{id}} = E_{\mathrm{id}}(\boldsymbol{F}_X, \boldsymbol{m}_X), \qquad \boldsymbol{Z}^X_{\mathrm{gar}} = E_{\mathrm{gar}}(\boldsymbol{F}_X, \boldsymbol{m}_X). \tag{5}$$

The identity representation captures body shape and subject-specific non-clothing information, while the garment representation captures clothing appearance, geometry, and coverage. We impose a stronger bottleneck on the identity branch than on the garment branch. Specifically, the latent shapes are $\boldsymbol{Z}^X_{\mathrm{id}} \in \mathbb{R}^{64\times128}$ and $\boldsymbol{Z}^X_{\mathrm{gar}} \in \mathbb{R}^{256\times128}$. Thus, the garment branch uses four times as many latent tokens, while both branches use a feature dimension of 128. This capacity asymmetry discourages garment information from leaking into the identity representation while preserving the variation required to reconstruct the garment. After decoupling, the reference garment must conform to the source body rather than reproduce the reference subject. We therefore decode a source-conditioned person representation from $\boldsymbol{Z}^A_{\mathrm{id}}$, and decode the garment representation from $\boldsymbol{Z}^B_{\mathrm{gar}}$ conditioned on the same source identity,

$$\boldsymbol{F}_{\mathrm{person}} = D_{\mathrm{person}}\left(\boldsymbol{Z}^A_{\mathrm{id}}\right), \quad \boldsymbol{F}_{\mathrm{gar}} = D_{\mathrm{gar}}\left(\boldsymbol{Z}^B_{\mathrm{gar}}, \boldsymbol{Z}^A_{\mathrm{id}}\right). \tag{6}$$

This conditioning adapts the reference garment to the source body while retaining its appearance and geometry. The garment coverage must also follow the reference outfit, particularly when the source and reference garments have different extents. We therefore use the target routing mask $\boldsymbol{m}_B$ to compose the person and garment representations

$$\boldsymbol{F}_{A\leftarrow B} = (1 - \boldsymbol{m}_B) \odot \boldsymbol{F}_{\mathrm{person}} + \boldsymbol{m}_B \odot \boldsymbol{F}_{\mathrm{gar}}, \tag{7}$$

where the routing mask is broadcast along the feature dimension. This composition preserves source-conditioned features outside the target garment support and inserts target garment information within

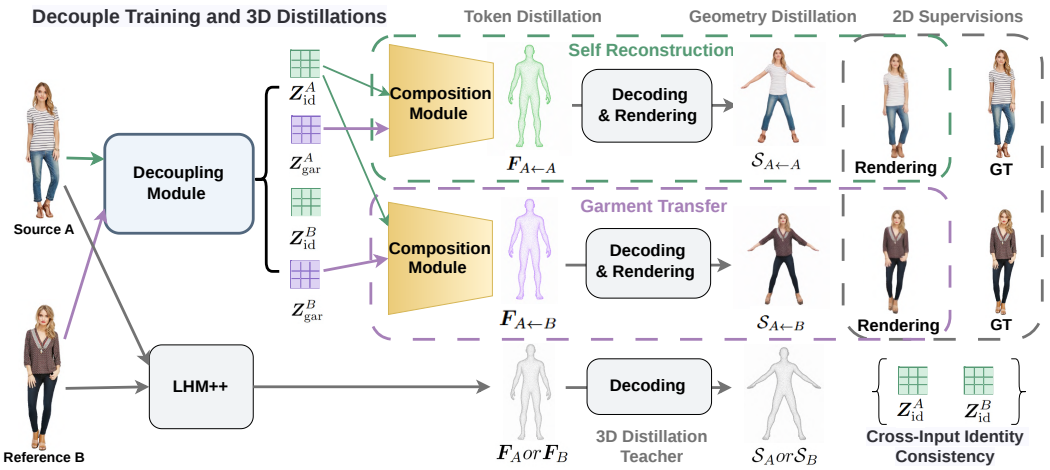


Figure 2: **Decoupling training and 3D distillation.** The decoupling module extracts identity and garment representations from **same-identity, different-clothing** pairs, which are then recomposed for self-reconstruction and cross-outfit garment transfer. Training combines token and geometry distillation from the frozen 3D teacher with 2D rendering supervision, while cross-input identity consistency encourages stable identity representations across different garments.

it. At inference, the resulting canonical token representation is rendered with the source pose $\mathcal{P}_A$ and produces both Gaussian and image outputs through the LHM++ decoders

$$\mathcal{S}_{A\leftarrow B} = \mathcal{G}(\boldsymbol{F}_{A\leftarrow B}), \qquad \hat{I}_{A\leftarrow B} = \mathcal{R}(\boldsymbol{F}_{A\leftarrow B}; \mathcal{S}_{A\leftarrow B}, \mathcal{P}_A). \tag{8}$$

Since $\mathcal{G}$ and $\mathcal{R}$ remain fixed, the transfer module must produce edited features compatible with the pretrained Gaussian human representation. Decoupling also makes the garment representation independently manipulable. Given two garment codes, intermediate outfits can be obtained through

$$\boldsymbol{Z}^{\alpha}_{\text{gar}} = (1-\alpha)\boldsymbol{Z}^{A}_{\text{gar}} + \alpha \boldsymbol{Z}^{B}_{\text{gar}}, \qquad \alpha \in [0,1], \tag{9}$$

while keeping the source identity representation fixed. This enables continuous garment manipulation within the same feed-forward transfer framework.

3.3 Decoupling Training Objectives

We train WearAnything using pairs (I_A, I_B) depicting the same identity in different outfits. This pairing provides direct supervision for cross-outfit transfer because the identity is shared while clothing changes between the two observations. For each pair, we optimize a self reconstruction path and a garment transfer path. The self path reconstructs an input from its own identity and garment representations, while the transfer path combines the source identity with the reference garment,

$$\boldsymbol{F}_{A\leftarrow A} = \text{Compose}\left(\boldsymbol{Z}^{A}_{\text{id}}, \boldsymbol{Z}^{A}_{\text{gar}}, \boldsymbol{m}_A\right), \quad \boldsymbol{F}_{A\leftarrow B} = \text{Compose}\left(\boldsymbol{Z}^{A}_{\text{id}}, \boldsymbol{Z}^{B}_{\text{gar}}, \boldsymbol{m}_B\right). \tag{10}$$

The self path stabilizes both reconstruction and the decoupling of identity from garment information, while the transfer path learns the desired cross-outfit composition.

3.3.1 3D Token and Geometry Distillation

Directly editing the canonical token representation can move it away from the feature distribution learned by LHM++, causing unstable Gaussian geometry. We therefore use the frozen LHM++ as a teacher in both the canonical token space and the explicit Gaussian attribute space. Because I_A and I_B depict the same person in different outfits, the frozen outputs $(\boldsymbol{F}_B, \mathcal{S}_B)$ provide direct 3D targets for the transfer result $(\boldsymbol{F}_{A\leftarrow B}, \mathcal{S}_{A\leftarrow B})$. Likewise, $(\boldsymbol{F}_A, \mathcal{S}_A)$ supervise self reconstruction.

For a soft canonical region weight $\boldsymbol{w}$, we define the normalized token distance and the corresponding Gaussian distance as:

$$D_{\text{tok}}(\boldsymbol{F}, \boldsymbol{F}^*; \boldsymbol{w}) = \frac{\sum_n \boldsymbol{w}_n \|\boldsymbol{F}_n - \boldsymbol{F}^*_n\|_1}{\sum_n \boldsymbol{w}_n + \epsilon}, \quad D_{\text{GS}}(\mathcal{S}, \mathcal{S}^*; \boldsymbol{w}) = \frac{\sum_n \boldsymbol{w}_n d_{\text{GS}}(\mathcal{S}_n, \mathcal{S}^*_n)}{\sum_n \boldsymbol{w}_n + \epsilon}, \tag{11}$$

where the per-Gaussian distance d_{GS} combines opacity, scale, position offset, and rotation terms with attribute-specific weights. The opacity and position offset terms use ℓ_1 distance, and the scale term uses ℓ_1 distance in log space. Gaussian appearance attributes are excluded. We omit $\boldsymbol{w}$ from D_{tok} and D_{GS} when supervision is global. The self reconstruction loss distills both representations:

$$\mathcal{L}_{\mathrm{self}}^{\mathrm{3D}} = D_{\mathrm{tok}}(\boldsymbol{F}_{A\leftarrow A}, \boldsymbol{F}_A) + D_{\mathrm{GS}}(\mathcal{S}_{A\leftarrow A}, \mathcal{S}_A). \tag{12}$$

This objective ensures that decoupling and recomposition preserve the original canonical tokens and 3D Gaussian avatar. For garment transfer, we retain weak global supervision and place stronger supervision on regions affected by the garment change. From the routing masks, we define $\boldsymbol{w}^+ = \boldsymbol{m}_B$, $\boldsymbol{w}^- = \boldsymbol{m}_A \odot (1 - \boldsymbol{m}_B)$, and $\boldsymbol{w}^\Delta = 1 - (1 - \boldsymbol{m}_A) \odot (1 - \boldsymbol{m}_B)$, where $\boldsymbol{w}^+$ is the target garment region, $\boldsymbol{w}^-$ is the source garment region to remove, and $\boldsymbol{w}^\Delta$ is the union of source and target garment regions. The swap token loss and swap Gaussian loss are:

$$\mathcal{L}_{\mathrm{tok}}^{\mathrm{swap}} = D_{\mathrm{tok}}(\boldsymbol{F}_{A\leftarrow B}, \boldsymbol{F}_B) + D_{\mathrm{tok}}(\boldsymbol{F}_{A\leftarrow B}, \boldsymbol{F}_B; \boldsymbol{w}^+) + D_{\mathrm{tok}}(\boldsymbol{F}_{A\leftarrow B}, \boldsymbol{F}_B; \boldsymbol{w}^\Delta), \tag{13}$$

$$\mathcal{L}_{\mathrm{GS}}^{\mathrm{swap}} = D_{\mathrm{GS}}(\mathcal{S}_{A\leftarrow B}, \mathcal{S}_B) + D_{\mathrm{GS}}(\mathcal{S}_{A\leftarrow B}, \mathcal{S}_B; \boldsymbol{w}^+) + D_{\mathrm{GS}}(\mathcal{S}_{A\leftarrow B}, \mathcal{S}_B; \boldsymbol{w}^-). \tag{14}$$

The weak global terms prevent the overall representation from drifting, while strong supervision on w^+ aligns the target garment features, geometry, and coverage in both spaces. In token space, w^Δ further enforces semantic consistency over the entire edited region. In Gaussian space, w^- is used instead to explicitly remove residual source-garment geometry, avoiding redundant supervision on w^+ and allowing garment insertion and removal to be controlled separately. The complete 3D distillation objective is:

$$\mathcal{L}_{\mathrm{3D}} = \mathcal{L}_{\mathrm{self}}^{\mathrm{3D}} + \lambda_{\mathrm{swap}} \left(\mathcal{L}_{\mathrm{tok}}^{\mathrm{swap}} + \mathcal{L}_{\mathrm{GS}}^{\mathrm{swap}} \right). \tag{15}$$

3.3.2 2D Rendering Supervision

We complement 3D distillation with rendering space supervision for garment appearance and visible silhouette. During training, the self reconstruction $\boldsymbol{F}_{A\leftarrow A}$ is rendered with the source pose $\mathcal{P}_A$ and supervised by I_A, while the transfer result $\boldsymbol{F}_{A\leftarrow B}$ is rendered with the reference pose $\mathcal{P}_B$ in the camera of I_B and supervised by I_B. Rendering under $\mathcal{P}_B$ aligns the prediction with the reference observation, and the shared identity makes I_B a direct image target for garment transfer. For a rendered prediction $\hat{I}$ and its corresponding target I^*, we use the foreground RGB loss $\mathcal{L}_{\mathrm{rgb}} = \|\boldsymbol{M}_{\mathrm{fg}} \odot (\hat{I} - I^*)\|_1$ together with the perceptual loss Zhang et al. (2018) $\mathcal{L}_{\mathrm{perc}} = \mathrm{LPIPS}(\hat{I}, I^*)$.

For the transfer path, we additionally emphasize regions affected by the garment change. A local garment reconstruction term focuses the image loss on the transferred region, while a crop level perceptual term compares garment crops at a fixed resolution. These terms encourage preservation of target garment texture, color, and fine appearance details that may contribute only weakly to the global image loss. We supervise the rendered human silhouette using the mask loss $\mathcal{L}_\alpha = \|\hat{\boldsymbol{M}}_{\mathrm{fg}} - \boldsymbol{M}_{\mathrm{fg}}^*\|_1$.

The routing head is supervised independently using garment segmentation. We render the canonical routing vector $\boldsymbol{m}_X$ into the image plane and optimize it against the image plane garment mask $\boldsymbol{M}_X^*$:$\mathcal{L}_{\mathrm{route}} = \mathrm{BCE}\left(\hat{\boldsymbol{M}}_X, \boldsymbol{M}_X^*\right)$. The complete rendering objective is

$$\mathcal{L}_{\mathrm{2D}} = \lambda_{\mathrm{rgb}}\mathcal{L}_{\mathrm{rgb}} + \lambda_{\mathrm{perc}}\mathcal{L}_{\mathrm{perc}} + \lambda_{\mathrm{local}}\mathcal{L}_{\mathrm{local}} + \lambda_{\mathrm{crop}}\mathcal{L}_{\mathrm{crop}} + \lambda_\alpha\mathcal{L}_\alpha. \tag{16}$$

Applying the relevant terms to the self and transfer predictions gives $\mathcal{L}_{\mathrm{self}}^{\mathrm{2D}}$ and $\mathcal{L}_{\mathrm{swap}}^{\mathrm{2D}}$, respectively; the local and crop level terms are used only for the transfer path.

3.3.3 Total Objective and Gradient Isolation

For same identity pairs, we further encourage the identity representation to remain invariant to clothing: $\mathcal{L}_{\mathrm{id}} = 1 - \cos\left(\bar{\boldsymbol{Z}}_{\mathrm{id}}^A, \bar{\boldsymbol{Z}}_{\mathrm{id}}^B\right)$, where $\bar{\boldsymbol{Z}}_{\mathrm{id}}$ denotes the pooled identity representation. The total training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{3D}} + \mathcal{L}_{\mathrm{self}}^{\mathrm{2D}} + \lambda_{\mathrm{swap}}\mathcal{L}_{\mathrm{swap}}^{\mathrm{2D}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{route}}\mathcal{L}_{\mathrm{route}}. \tag{17}$$

Here, λ_{swap} weights the transfer path once in both the 3D and 2D objectives. We use stronger region aware supervision for transfer and global reconstruction supervision for self reconstruction.

Finally, we isolate gradients to the routing head. The routing prediction receives gradients only from $\mathcal{L}_{\mathrm{route}}$; during training, $\boldsymbol{m}_A$ and $\boldsymbol{m}_B$ are detached wherever they are used for feature composition and region weighting. This prevents reconstruction losses from modifying garment coverage to reduce their own objectives and keeps the routing field aligned with semantic clothing regions.



Figure 3: **Qualitative comparison with existing virtual try-on methods.** WearAnything matches the category and spatial coverage of the reference garment while preserving the source identity and body shape.

4 EXPERIMENTS

Dataset. DeepFashion-MultiModal Jiang et al. (2022) is used to construct same-identity, different-garment training pairs. Since the original dataset does not provide cross-outfit identity correspondences, we perform face-based identity matching and retain subjects observed in multiple outfits. The resulting dataset contains 592 identities and 255750 unordered cross-outfit image pairs for training our garment transfer model.

Implementation Details. The model is implemented in PyTorch and trained on 3 NVIDIA H100 GPUs. The pretrained LHM++ reconstructor is kept frozen, and our garment-transfer module is trained, which contains 4.12M trainable parameters. Training is conducted for 40,000 iterations using bfloat16 precision, with a per-GPU batch size of 3 taking approximately 72 hours. Additional training and loss details are provided in the Appendix.

4.1 EVALUATIONS

We evaluate WearAnything on three out-of-domain datasets: StreetTryOn Cui et al. (2025), MVHumanNet Xiong et al. (2024) and In-the-Wild datasets. We compare our method against state-of-the-art 3D virtual try-on methods, including M3D-VTON Zhao et al. (2021), GS-VTON Cao et al. (2026), and VTON360 He et al. (2025). We also include two recent 2D VTON methods, OOTDiffusion Xu et al. (2025) and OneModelForAll Cheng et al. (2025), which are designed for 2D garment transfer but can be extended to 3D by adding a pretrained 3D reconstructor LHM++ Qiu et al. (2025b). All evaluations use a single RTX 3090 GPU.

Table 1: Quantitative comparison of garment transfer on StreetTryOn Cui et al. (2025) and MVHumanNet Xiong et al. (2024). Baselines include M3D-VTON Zhao et al. (2021), GS-VTON Cao et al. (2026), VTON360 He et al. (2025), and two 2D VTON plus LHM++ reconstruction pipelines based on OOTDiffusion Xu et al. (2025) and OneModelForAll Cheng et al. (2025).

Method	#Stages↓	StreetTryOn (single-view)				MVHumanNet (multi-view)			
		PSNR↑	SSIM↑	LPIPS↓	Time (s)↓	PSNR↑	SSIM↑	LPIPS↓	Time (s)↓
M3D-VTON	4	16.74	0.5364	0.2149	29.3	16.28	0.5169	0.2237	51.1
GS-VTON	4	17.93	0.6524	0.1862	1283.6	17.29	0.6341	0.1972	1832.6
VTON360	3	19.53	0.7548	0.1492	354.8	16.87	0.5928	0.2074	401.6
OOTDiffusion(+ 3D)	4+1	18.32	0.6921	0.1614	60.3	17.56	0.6348	0.1782	184.9
OneModelForAll(+ 3D)	3+1	20.65	0.8249	0.1293	88.7	19.35	0.7934	0.1348	236.2
Ours	**1**	**23.65**	**0.8567**	**0.1034**	**5.4**	**22.91**	**0.8479**	**0.1230**	**5.7**

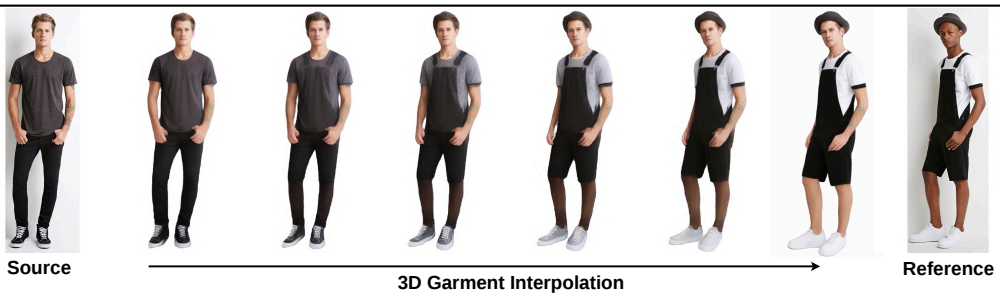


Figure 4: **Continuous 3D garment interpolation.** WearAnything smoothly transitions between source and reference garments in latent space while preserving the source identity and body shape.

Qualitative Evaluation. Qualitative comparisons are shown in Figure 3. The examples cover substantial changes in garment category and spatial support. M3D-VTON and GS-VTON mainly modify appearance within the source garment support, often retaining source trousers, skirts, or shorts when the target garment has a different extent. VTON360 and OOTDiffusion can expand the edited region but frequently produce inaccurate garment length, color, or boundaries, together with visible artifacts. OneModelForAll more closely matches the target garment category and coverage, but often alters facial appearance and body proportions, weakening source identity preservation. In contrast, WearAnything consistently reproduces the target garment type and coverage while retaining the source face and body shape. These results demonstrate the advantage of learning a garment latent space and adapting feature composition to the target garment support for cross-category transfer.

Quantitative Evaluation. For quantitative evaluation, we construct same-identity, different-clothing pairs (I_A, I_B) from StreetTryOn Cui et al. (2025) and MVHumanNet Xiong et al. (2024), using the former for single-view evaluation and the latter for multi-view evaluation. We retain pairs with clear clothing changes while preserving the subject identity. For each pair, the transferred avatar is rendered using the pose and camera of the reference image I_B, which serves as the ground truth for PSNR, SSIM Wang et al. (2004), and LPIPS Zhang et al. (2018). This protocol provides pose aligned image space comparisons with the same subject wearing the target garment. Further details on evaluation set construction are provided in the Appendix.

Table 1 reports reconstruction quality, end-to-end runtime, and the number of processing stages. Runtime and stage count are measured from the raw source and reference images to the final 3D avatar, including all required preprocessing and intermediate generation, optimization, and reconstruction steps. WearAnything achieves the best reconstruction quality on both benchmarks while completing garment transfer and 3D avatar generation. In the single-view setting, existing methods rely on diffusion inpainting to synthesize the target garment before 3D reconstruction. Inpainting errors in garment shape, texture, or identity are therefore carried into the reconstructed avatar. The multi-view setting introduces an additional challenge because only sparse input views are available. Beyond errors inherited from inpainting, competing methods must infer consistent geometry and appearance in unobserved views, which is difficult for pipelines designed around dense multi-view observations. WearAnything instead transfers the garment directly within a canonical 3D representation, producing consistent geometry and appearance from sparse inputs while avoiding repeated editing and reconstruction.

Table 2: **Ablation study of WearAnything.**

Ablations	PSNR ↑	SSIM ↑	LPIPS ↓	Mask IoU ↑	Dice ↑	ArcFace Sim. ↑
w/o Decoupling	16.53	0.5920	0.3652	0.6637	0.7022	0.4609
w/o Garment Routing	17.65	0.6533	0.2518	-	-	0.5132
w/o 3D Losses	20.31	0.7653	0.1636	0.7924	0.8215	0.6433
w/o 2D Losses	19.56	0.7192	0.1982	0.7651	0.7792	0.5756
w/o Gradient Isolation	22.95	0.8093	0.1168	0.8235	0.8629	0.6926
Full Model	**23.65**	**0.8567**	**0.1034**	**0.8918**	**0.9420**	**0.7112**

4.2 Garment Interpolation

Given the continuous structure of the learned garment latent space, WearAnything supports smooth interpolation between different outfits. As shown in Figure 4, interpolating between two garment representations produces gradual changes in clothing appearance and geometry while preserving the source identity and body shape. This result indicates that the learned garment space captures meaningful garment variation and enables controllable editing beyond discrete garment transfer.

4.3 Ablation Study

Table 2 evaluates identity and garment decoupling, garment routing, 3D and 2D supervision, and gradient isolation. In addition to image reconstruction metrics, we measure garment-mask accuracy using Intersection over Union (IoU) and the Dice coefficient Taha & Hanbury (2015), and identity consistency using the cosine similarity between ℓ_2-normalized ArcFace embeddings Deng et al. (2019). The full model performs best across all metrics. Removing decoupling causes the largest overall degradation and sharply reduces identity similarity, showing that direct feature composition transfers reference identity information and corrupts non-garment regions. Removing garment routing prevents mask prediction, making IoU and Dice inapplicable, while also degrading reconstruction and identity preservation because garment features cannot be localized reliably. Removing either 3D or 2D supervision reduces mask accuracy and identity consistency together with image quality, confirming that canonical geometry constraints and rendered appearance supervision are complementary. Finally, removing gradient isolation lowers Mask IoU, Dice, and ArcFace similarity despite a smaller change in reconstruction metrics, supporting the use of detached routing masks to keep garment support semantically aligned and prevent reconstruction objectives from affecting identity preservation.

5 Discussion and Conclusions

Discussion. Although WearAnything demonstrates promising results in 3D garment transfer, there are still some limitations to address. First, the model's performance may degrade when dealing with extreme poses or occlusions, as the pretrained reconstructor may struggle to accurately capture the underlying human geometry. Second, while our method supports cross-category garment transfer, it may still face challenges when transferring garments with highly complex structures or intricate details. Future work could explore incorporating more advanced geometric priors or leveraging additional multi-view data to enhance the robustness of the model.

Conclusion. In this paper, we presented WearAnything, a feed-forward framework for efficient and controllable 3D garment transfer in Gaussian human representations. By operating directly in the canonical token space of a pretrained human reconstructor, WearAnything avoids multi-stage reconstruction and subject-specific optimization while enabling fast end-to-end inference. A lightweight identity and garment decoupling module with routing-guided composition supports localized cross-category garment transfer, while our decoupling training strategy with 3D feature and geometry distillation preserves the structural validity and view-consistent geometry of the transferred avatar. Extensive experiments across multiple datasets demonstrate state-of-the-art garment transfer quality, strong identity preservation, and substantially improved efficiency. In addition, the learned garment latent space supports continuous outfit interpolation, highlighting the potential of canonical 3D representations for flexible and controllable avatar editing.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, to draft portions of the manuscript, and to assist with retrieval, discovery of related work and draft that section. We did not use generative AI tools for research ideation or execution, synthetic dataset generation, or proving mathematical claims. All AI-assisted text was reviewed and revised by the authors, and retrieved literature was manually verified against the original sources before inclusion. We take full responsibility for the final content of this work, including all text, claims, citations, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to support reproducibility by documenting the model design, training procedure, implementation settings, dataset construction, evaluation protocol, and ablation setup throughout the main paper and Appendix. Additional experimental details are provided in the supplementary material. Subject to company approval, we also plan to release the implementation code and relevant configuration files to further facilitate reproduction of our results.

REFERENCES

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

Yukang Cao, Masoud Hadi, Liang Pan, and Ziwei Liu. Gs-vton: Controllable 3d virtual try-on with gaussian splatting. *International Journal of Computer Vision*, 134(5):215, 2026.

Haodong Chen, Yongle Huang, Haojian Huang, Xiangsheng Ge, and Dian Shao. Gaussianvton: 3d human virtual try-on via multi-stage gaussian splatting editing with image prompting. *arXiv preprint arXiv:2405.07472*, 2024.

Zhiyi Chen, Jie Song, Hsuan-I Ho, Manuel Kaufmann, Tianjian Jiang, and Chen Guo. Gaussian wardrobe: Compositional 3d gaussian avatars for free-form virtual try-on. In *2026 International Conference on 3D Vision (3DV)*, pp. 739–749. IEEE, 2026.

Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One model for all: Low-level task interaction is a key to task-agnostic image fusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28102–28112. IEEE, 2025.

Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, Chang Liu, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1414–1423, 2025. doi: 10.1109/WACV61041.2025.00145.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. doi: 10.1109/CVPR.2019.00482.

Chenghu Du, Feng Yu, Minghua Jiang, Yaxin Zhao, Xiong Wei, Tao Peng, and Xinrong Hu. Realistic monocular-to-3d virtual try-on via multi-scale characteristics capture. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1735–1739. IEEE, 2022.

Zijian He, Yuwei Ning, Yipeng Qin, Guangrun Wang, Sibei Yang, Liang Lin, and Guanbin Li. Vton 360: High-fidelity virtual try-on from any viewing direction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26388–26398, 2025. doi: 10.1109/CVPR52734.2025.02457.

Xinrong Hu, Cheng Zheng, Junjie Huang, Ruiqi Luo, Junping Liu, and Tao Peng. Cloth texture preserving image-based 3d virtual try-on. *The Visual Computer*, 39(8):3347–3357, 2023.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4): 1–11, 2022. doi: 10.1145/3528223.3530104.

Daisheng Jin, Shuyun Wang, and Ying He. Forwardrobe: Garment-aware gaussian avatars from a single image. *arXiv preprint arXiv:2607.29106*, 2026.

Jiahui Lei, Srinath Sridhar, Paul Guerrero, Minhyuk Sung, Niloy Mitra, and Leonidas J Guibas. Pix2surf: Learning parametric 3d surface models of objects from images. In *European Conference on Computer Vision*, pp. 121–138. Springer, 2020.

Min Li, Ying Zhang, Yanping Tian, Junyin Wang, and Chenghu Du. Humod: High-quality human modeling from monocular virtual try-on image. *IEEE Transactions on Consumer Electronics*, 2024.

Siyou Lin, Zhe Li, Zhaoqi Su, Zerong Zheng, Hongwen Zhang, and Yebin Liu. Layga: Layered gaussian avatars for animatable clothing transfer. In *ACM SIGGRAPH 2024 Conference Papers*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14184–14194, 2025a. doi: 10.1109/ICCV51701.2025.01316.

Lingteng Qiu, Peihao Li, Heyuan Li, Qi Zuo, Xiaodong Gu, Yuan Dong, Weihao Yuan, Rui Peng, Siyu Zhu, Xiaoguang Han, Guanying Chen, and Zilong Dong. Lhm++: An efficient large human reconstruction model for pose-free images to 3d. *arXiv preprint arXiv:2503.10625*, 2025b.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, 2015. doi: 10.1186/s12880-015-0068-x.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Zhenyu Xie, Haoye Dong, Yufei Gao, Zehua Ma, and Xiaodan Liang. Dreamvton: Customizing 3d virtual try-on with personalized diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10784–10793, 2024.

Zhenyu Xie, Fuwei Zhao, Jun Zheng, Xin Dong, Feida Zhu, and Xiaodan Liang. Monocular-to-3d virtual try-on with generative semantic articulated fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):9718–9730, 2025. doi: 10.1109/TPAMI.2025.3591072.

Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, Shuguang Cui, and Xiaoguang Han. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19801–19811, June 2024.

Yinghan Xu and John Dingliana. Layergs: Decomposition and inpainting of layered 3d human avatars via 2d gaussian splatting. In *International Conference on Pattern Recognition*, pp. 270–284. Springer, 2026.

Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8996–9004, 2025.

Feng Zhang, Maochun Zhang, Zhenming Chen, Hao Feng, Biao Guo, Junyan Chen, Yao Lu, and Ming Jiang. Anyview vton: Consistent 3d virtual try-on with view-conditioned diffusion model. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 482–496. Springer, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13219–13229, 2021. doi: 10.1109/ICCV48922.2021.01299.

Xiaojing Zhong, Yukun Su, Zhonghua Wu, Guosheng Lin, and Qingyao Wu. Di-net: Decomposed implicit garment transfer network for digital clothed 3d human. *Pattern Recognition*, pp. 112574, 2025a.

Xiaojing Zhong, Zhonghua Wu, Xiaofeng Yang, Guosheng Lin, and Qingyao Wu. Ipvton: Image-based 3d virtual try-on with image prompt adapter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10671–10679, 2025b.

A Appendix

A.1 More Related Works

A.1.1 Human Reconstruction and Editing

Recent human reconstruction methods increasingly adopt 3D Gaussian representations to model detailed and animatable human avatars. Feed-forward approaches such as LHM Qiu et al. (2025a) further demonstrate that a single image can be mapped directly to a structured Gaussian human representation with high efficiency. These models provide strong geometric and appearance priors, but their representations are primarily optimized for faithful reconstruction. As a result, identity, body structure, and clothing information are often encoded jointly, making direct garment-level manipulation difficult.

A related line of work focuses on improving avatar editability by explicitly separating garments from the human body. LayGA Lin et al. (2024) reconstructs layered Gaussian avatars for independent body and clothing manipulation, while Gaussian Wardrobe Chen et al. (2026) represents garments as reusable components for compositional avatar editing. LayerGS Xu & Dingliana (2026) further decomposes the human representation into body and garment layers to support localized editing, and Forwardrobe Jin et al. (2026) explores garment-aware Gaussian reconstruction with separable clothing representations. These methods improve clothing-level controllability but mainly focus on constructing editable avatar representations. In contrast, WearAnything performs garment transfer directly in the canonical representation of a pretrained human reconstructor, enabling efficient feed-forward virtual try-on without constructing a separate layered avatar for each subject.

A.2 Model Details

Our garment transfer module operates on the canonical tokens extracted by the frozen reconstruction backbone. Each input image is represented by N canonical tokens with a feature dimension of 128. Only the following modules are trained.

Garment Routing. The routing head is a three-layer MLP shared by the source and reference inputs. It consists of LayerNorm, two 256-dimensional hidden layers with GELU activations, and a scalar output layer. It takes the canonical features of shape $N \times 128$ as input and predicts one garment logit for each token.

The routing prediction is further conditioned on the input garment mask. We project the 2D mask onto the canonical Gaussian centers and obtain a mask prior and a projection-validity value for every token. A lightweight refinement network takes the 128-dimensional token feature, a 16-dimensional canonical coordinate embedding, the mask prior, and its validity as input. The features are processed at 64 dimensions and aggregated over eight fixed nearest neighbors on the canonical lattice. The network outputs one residual logit per token. Combining the MLP prediction, the refinement residual, and the projected mask prior produces the final soft routing mask of shape $N \times 1$.

Identity and Garment Encoders. The routing mask separates the canonical features into identity and garment streams. Masked regions are replaced with learned tokens rather than zeros. Both streams are encoded using four-layer Perceiver-style cross-attention Jaegle et al. (2021) networks with eight attention heads. The identity encoder takes an $N \times 128$ routed feature field and outputs 64 identity latent tokens of dimension 128. The garment encoder has the same architecture but outputs 256 garment latent tokens, providing additional capacity for garment appearance and local details.

Canonical Query Embedding. A content-free query is generated from the canonical 3D coordinates. The coordinates are first converted into a 48-dimensional Fourier positional encoding Tancik et al. (2020) and concatenated with the original 3D coordinates. A linear layer and LayerNorm map the result to an $N \times 128$ query field. Because this query contains no image features, it prevents appearance leakage from masked regions.

Identity Decoder. The identity decoder contains four cross-attention blocks with eight attention heads. It takes the sanitized canonical query of shape $N \times 128$ and the 64 identity latent tokens as input. Its output is an identity-preserving canonical feature field of shape $N \times 128$.

Garment Decoder. The garment decoder contains two consecutive groups of cross-attention blocks. The first four blocks attend to the 256 reference garment latent tokens, and the following four blocks attend to the 64 source identity latent tokens. It takes the sanitized garment query as input and outputs a garment-conditioned feature field of shape $N \times 128$.

Feature Composition. The identity and garment decoder outputs are blended using the reference routing mask. Non-garment tokens are selected from the identity branch, while garment tokens are selected from the garment branch. The resulting $N \times 128$ canonical feature field is passed to the frozen Gaussian decoder and renderer, which produce the transferred 3D Gaussians and rendered RGB image.

A.3 Dataset Preprocessing and Training Details

Dataset Preprocessing. DeepFashion-MultiModal Jiang et al. (2022), StreetTryOn Cui et al. (2025), and MVHumanNet Xiong et al. (2024) provide RGB images together with corresponding segmentation masks. We preprocess each dataset in three stages. First, we group sequences belonging to the same subject using face based identity matching with InsightFace **?**. Second, we apply MediaPipe Pose Landmarker Bazarevsky et al. (2020) to retain full body images in which the head and both feet are visible. Third, we crop each subject according to the detected pose landmarks, preserve the original aspect ratio, and resize the resulting image to 504×840.

Evaluation Set Construction. We randomly select 50 same-identity, different-clothing pairs from StreetTryOn and 50 pairs from MVHumanNet. To prevent data leakage, we verify that the evaluation

identities and images are fully disjoint from the training set, with no subject or image overlap. We also manually inspect every selected pair to validate the face-based identity match and retain only pairs that clearly depict the same person in different clothing. For the single-view StreetTryOn benchmark, we use frontal source and reference images as inputs and evaluate the transferred result only in the frontal target view. For the multi-view MVHumanNet benchmark, the front, side, and back views are provided as inputs, and eight additional views excluded from the inputs are used for novel-view evaluation.

Evaluation Details. In both settings, the ground-truth images depict the same subject wearing the reference garment, and predictions are rendered using the corresponding target poses and cameras before computing PSNR, SSIM, and LPIPS. All methods are evaluated on exactly the same fixed pairs, enabling direct paired comparisons. The MVHumanNet evaluation contains 400 novel-view comparisons in total. We first average the eight view-level scores within each pair and then average across the 50 pairs, ensuring equal contribution from every pair and preventing any individual subject or view from dominating the results. AnyView VTON Zhang et al. (2025), G3D-VTON Xie et al. (2025), and IPVTON Zhong et al. (2025b) are not included because their official implementations and pretrained models are not publicly available. We therefore cannot evaluate them reproducibly under the same datasets, inputs, and metric protocol, and avoid reporting results from unofficial reimplementations.

Training Details. We keep the reconstruction backbone and Gaussian renderer frozen and optimize only the proposed garment transfer module. At inference, when an input segmentation mask is unavailable, we automatically obtain the required mask using a human parsing model with a SegFormer-B2 backbone Xie et al. (2021). Training uses AdamW Loshchilov & Hutter (2017) with a learning rate of 1×10^{-4} and a weight decay of 5×10^{-4}. We train for 40 epochs with 1,000 iterations per epoch and a per-GPU batch size of 3. The learning rate follows a cosine schedule with 1,000 warm-up iterations. The trainable transfer module uses BF16 mixed precision, while frozen reference features and loss computation remain in FP32. Training contains the self reconstruction and garment transfer paths defined in the main paper. The transfer weight is set to $\lambda_{\text{swap}} = 0.5$ for the first 2,000 iterations and to 1.0 afterward.

3D Distillation. In $\mathcal{L}^{\text{3D}}_{\text{self}}$, the global D_{tok} and D_{GS} terms use weights of 0.05 and 1.0, respectively. For $\mathcal{L}^{\text{swap}}_{\text{tok}}$, the global, target garment $\boldsymbol{w}^{+}$, and garment change $\boldsymbol{w}^{\Delta}$ terms use weights of 0.05, 0.30, and 0.30. For $\mathcal{L}^{\text{swap}}_{\text{GS}}$, the corresponding global, target garment $\boldsymbol{w}^{+}$, and garment removal $\boldsymbol{w}^{-}$ terms use weights of 0.10, 1.0, and 0.50. Within d_{GS}, the opacity, scale, position offset, and rotation weights are $(1.0, 1.0, 1.0, 1.0)$ for self reconstruction and global transfer, $(0.7, 0.5, 0.2, 0.1)$ on $\boldsymbol{w}^{+}$, and $(0.2, 0.2, 0.1, 0.1)$ on $\boldsymbol{w}^{-}$. Gaussian appearance attributes are excluded from direct supervision.

2D Rendering Supervision. For both $\mathcal{L}^{\text{2D}}_{\text{self}}$ and $\mathcal{L}^{\text{2D}}_{\text{swap}}$, we set $\lambda_{\text{rgb}} = 0.5$ and $\lambda_{\text{perc}} = 0.5$. The transfer path additionally uses $\lambda_{\text{local}} = 1.0$ and $\lambda_{\text{crop}} = 0.1$, while the silhouette term uses $\lambda_{\alpha} = 0.1$. All RGB and perceptual losses supervise the frozen DPT output; direct appearance supervision on the raw Gaussian rendering is disabled.

Routing and Identity Supervision. The routing and identity terms in the total objective use $\lambda_{\text{route}} = 0.18$ and $\lambda_{\text{id}} = 0.5$, respectively. The routing mask is rendered into image space for $\mathcal{L}_{\text{route}}$, while $\mathcal{L}_{\text{id}}$ enforces consistency between the source and reference identity latents.

Loss Scheduling. Global DPT RGB and LPIPS losses are linearly activated during the first 1,000 iterations. Garment-region token and Gaussian losses, together with the local RGB loss, start at iteration 1,000 and reach their full weights after another 2,000 iterations. The garment-crop LPIPS loss starts at iteration 3,000 and is linearly activated over 500 iterations. For robust mask conditioning, the input garment mask is kept unchanged with probability 0.8, morphologically perturbed with probability 0.1, and removed with probability 0.1.

Table 3: **Cross-identity ArcFace Deng et al. (2019) similarity.**

Datasets	In-The-Wild	StreetTryOn	MVHumanNet
Ours	0.6961	0.7112	0.6398
LHM++ (Recon. Reference)	0.7059	0.7260	0.6729

A.4 More Experiments

A.4.1 More Multi-view Evaluations

Figure 5 evaluates garment transfer when only sparse reference views are available while assessing the resulting avatars from multiple viewpoints. Even with front and back inputs, M3D-VTON Zhao et al. (2021) and GS-VTON Cao et al. (2026) produce incomplete garment transfer, visible artifacts, and inconsistent appearance across views; these problems become more severe with a single reference view. VTON360 relies on dense multi-view observations for its 3D lifting stage and fails to construct a valid 3D result in all four sparse-view settings. We therefore show only its edited 2D views to distinguish successful image editing from failed 3D reconstruction. In contrast, WearAnything transfers the target garment directly in the canonical representation and produces coherent front and back views from both two-view and single-view inputs. The results show that the learned garment representation captures sufficient appearance and geometry to recover unobserved regions without dense view coverage while preserving the source identity and body shape.

A.4.2 Cross-identity Evaluations

Existing datasets do not provide cross-identity ground-truth images showing a source subject wearing a garment taken from a different reference identity. Consequently, full-reference metrics such as PSNR, SSIM, and LPIPS cannot be computed for cross-identity transfer. We instead evaluate identity preservation using the cosine similarity between ℓ_2-normalized ArcFace embeddings of the source image and the transferred rendering. For each dataset, the source and reference are selected from different identities, with the reference contributing only the target garment. A higher similarity indicates that the transferred result remains closer to the source identity despite the cross-identity garment input. Table 3 shows that WearAnything remains close to the LHM++ reconstruction reference across all three datasets. This small gap is notable because WearAnything simultaneously transfers a garment from a different identity, whereas the reconstruction reference does not perform cross-identity composition. The result indicates that the garment representation and feature composition introduce little reference-identity leakage and preserve the source identity effectively. The moderately larger gap on MVHumanNet is expected because identity embeddings are measured from novel-view renderings, where changes in viewpoint and facial visibility make face matching more challenging.



Figure 5: **Sparse-view garment transfer with multi-view evaluation.** The first two rows use front and back reference views, while the rest use a single reference view. For VTON360 He et al. (2025), the crosses indicate failed 3D lifting, and the displayed images are its 2D edited results before lifting.