

AvatarMix: Identity-Preserving Cross-Avatar Composition for Outfit Personalization

Supplementary Material

Zhaorong Wang, Yoshihiro Kanamori, Yuki Endo
University of Tsukuba

zhaorong.wang1997@gmail.com, {kanamori, endo}@cs.tsukuba.ac.jp

A. Task Setting and Paradigm Comparison

To clarify the task setting and our position in the design space, we summarize three dominant paradigms for avatar outfit editing and personalization in Tab. 2: 2D-to-3D virtual try-on, layered 3D garment modeling, and explicit 3D composition on Gaussian avatars. The table contrasts these paradigms by input conditions, use of generative models, and how clothes-body intersections are handled. AvatarMix belongs to the explicit 3D composition paradigm and limits diffusion to localized seam and artifact refinement conditioned on 3D-consistent rendered images, in contrast to artifact-prone garment inpainting in 2D VTON and collision-prone layered 3D garments.

B. Implementation and Evaluation Details

B.1 Skin Tone Transfer

To harmonize skin appearance between the user’s head and the model’s body, we operate in Lab color space with opacity-weighted statistics over Gaussian colors. Let $\{\mathbf{c}_i \in [0, 1]^3, \alpha_i\}_{i=1}^N$ denote RGB colors and opacities of either the user’s facial Gaussians or the model’s skin Gaussians, and let $\ell_i = f_{\text{Lab}}(\mathbf{c}_i)$ be the Lab conversion of the RGB colors. We compute the opacity-weighted mean μ and variance σ^2 as

$$\mu = \frac{\sum_{i=1}^N \alpha_i \ell_i}{\sum_{i=1}^N \alpha_i}, \quad \sigma^2 = \frac{\sum_{i=1}^N \alpha_i (\ell_i - \mu)^2}{\sum_{i=1}^N \alpha_i}. \quad (1)$$

We denote the user’s facial statistics as (μ^u, σ^u) and the model’s skin statistics as (μ^m, σ^m) . For each model’s skin color ℓ in Lab color space, we perform channel-wise affine transfer

$$\ell' = (\ell - \mu^m) \odot \frac{\sigma^u}{\sigma^m} + \mu^u, \quad (2)$$

where \odot denotes element-wise multiplication. The transformed colors ℓ' are then converted back to RGB and assigned to the corresponding skin Gaussians. This global, opacity-aware color transfer matches the model body’s skin tone to the user’s while preserving local shading and high-frequency detail.

B.2 Diffusion Refinement and GSReshape Implementation

Diffusion refinement. Both SeamFix and FullbodyFix are implemented and trained on top of the pretrained Diffix3D+ backbone [59], as described in the main paper. In addition, we attach rank-4 LoRA adapters to the VAE decoder and fine-tune the skip connections between the VAE encoder and decoder following Diffix3D. SeamFix is trained for 10 epochs and FullbodyFix for 5 epochs on approximately 19k multi-view double-swapped training samples generated from the THUMAN2.0 training subjects, using a batch size of 1. For SeamFix, we operate on a cropped square head-and-neck region that is resized to 512×512 during both training and testing. At test time we paste the refined crop back into the original image with a feathered blending boundary. For FullbodyFix, we crop a tight bounding box around the full human body, resize this crop to 488×896 pixels, and use this resolution during training and inference. All training is conducted on a single NVIDIA RTX A6000 Ada GPU; training SeamFix and FullbodyFix requires roughly 16 and 28 hours, respectively.

GSReshape optimization. Our GSReshape module builds on the intersection-free garment retargeting method of Huang *et al.* [29]; an overview of the full pipeline is shown in Fig. 6. Before retargeting, we optimize the SMPL-X skeleton vertices $X = \{x_k\}_{k=1}^{N_s}$ inside a signed distance field $\phi(\cdot)$ of the clothed body mesh in order to remove bone-mesh intersections while preserving bone lengths. When initializing the skeleton from the SMPL mesh, we attach each vertex to the bone with the largest linear blend skinning (LBS) weight instead of using nearest-distance assignment, which yields a more stable optimization and is visualized by the color-coded SMPL vertices in Fig. 6. For each bone (i, j) , we sample a set of points $\{p_s\}$ along the segment and define an inside penalty

$$E_{\text{inside}}(X) = \sum_s \max(0, \phi(p_s) + \delta)^2, \quad (3)$$

with margin $\delta = 0.1$, a bone-length regularizer

$$E_{\text{length}}(X) = \sum_{(i,j) \in E} (\|x_i - x_j\|^2 - L_{ij}^2)^2, \quad (4)$$

$$L_{ij}^2 = \|x_i^{(0)} - x_j^{(0)}\|^2, \quad (5)$$

Table 2. **Comparison of avatar editing paradigms.** We group related work into three paradigms and compare them by inputs, generative model usage, and clothes-body intersection handling. Unlike 2D-to-3D VTON methods that rely on garment inpainting and layered 3D garment approaches that require collision post-processing, AvatarMix composes two Gaussian avatars explicitly and applies diffusion as localized refinement on 3D-consistent rendered images, which improves view consistency and facial identity preservation, and is intersection-free by design.

Paradigm	Representative methods	Input conditions	Generative model usage	Clothes-body intersection handling
2D-to-3D VTON	VTON360 [25], GS-VTON [7], Gaussian-VTON [9]	User multi-view images + garment images	Full image generation with garment inpainting	N/A
Layered 3D garments	LayGA [38]	Two multiview videos	N/A	Post-processing for collision handling
Gaussian avatars 3D composition	AvatarMix (ours)	Two Gaussian avatars	Local/global refinement on 3D-consistent renderings	N/A

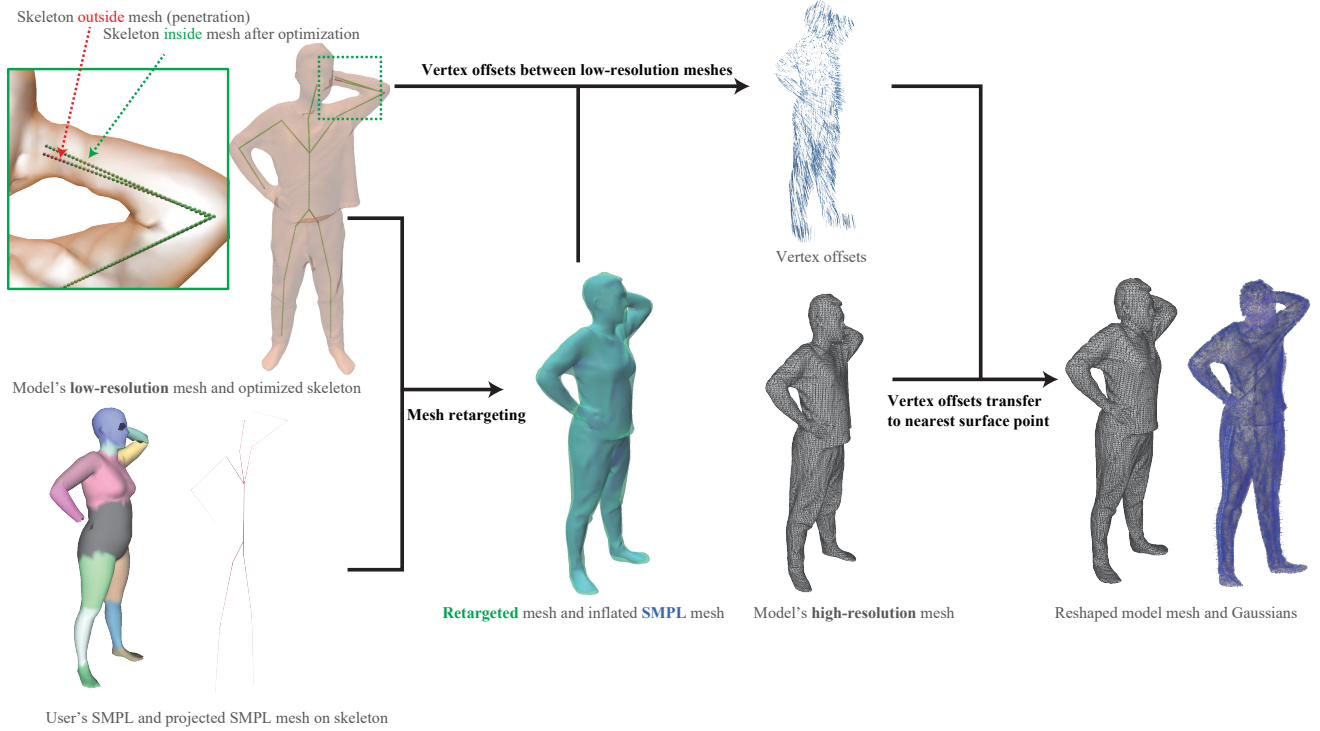


Figure 6. **GSReshape pipeline overview.** From left to right: starting from the model’s low-resolution clothed mesh (top left) and SMPL-X mesh (bottom left), we project the SMPL mesh to skeleton, inflating the SMPL mesh while jointly optimizing the clothed mesh, following the retargeting method of Huang *et al.* [29]. After retargeting, we compute vertex offsets between input and retargeted clothed mesh, and transfer these offsets to the original high-resolution clothed mesh via nearest-surface projection. The Gaussians defined on high-resolution mesh are updated as well.

and an root anchor regularizer

$$E_{\text{anchor}}(X) = \|x_r - x_r^{(0)}\|^2. \quad (6)$$

The skeleton optimization objective

$$E_{\text{pre}}(X) = w_{\text{inside}} E_{\text{inside}}(X) + w_{\text{len}} E_{\text{length}}(X) \quad (7)$$

$$+ w_{\text{anch}} E_{\text{anchor}}(X), \quad (8)$$

with $w_{\text{inside}} = 50.0$, $w_{\text{len}} = 5.0$, and $w_{\text{anch}} = 10.0$, is minimized with the same set of solvers as the method of Huang *et al.*, initialized from the original skeleton. We use 40 samples per bone, an SDF voxel size of 0.005, and robust SDF settings (flood-filled sign, hole closing with a 2-voxel radius, and capping of open boundaries), yielding an intersection-free and approximately rigid skeleton

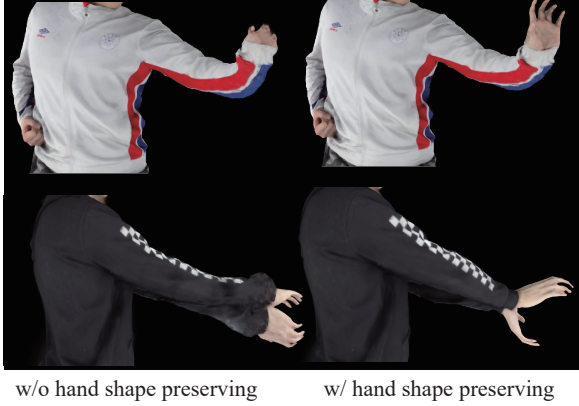


Figure 7. **Hand-aware skin tightness examples.** First row: high fit weight produces Gaussian artifacts (left) versus our hand shape preserving method (right). Second row: low fit weight creates glove-like hands (left) versus our approach (right). Our semantic weighting strategy achieves better balance between visual fidelity and robustness.

used in GSReshape. During the subsequent mesh retargeting, let $V_g = \{v_i\}$ and $X_A = \{x_j\}$ denote garment and avatar vertices. On vertices in the hand region we scale the SDF fit and similarity weights using $m_{\text{fit}}(v_i) = 0.01$ and $m_{\text{sim}}(v_i) = 2.0$, while keeping $m_{\text{fit}}(v_i) = 5$, $m_{\text{sim}}(v_i) = 1$ elsewhere. We also remove SMPL-X mesh vertices belonging to the hands from the avatar before SDF construction and continuation, which together help avoid glove-like inflation while preserving local hand shape. Fig. 7 illustrates the effectiveness of this hand-aware design with two examples: the first row compares high fit weight (causing Gaussian artifacts) with our hand shape preserving method, and the second row compares low fit weight (producing glove-like hands) with our approach, demonstrating that our semantic weighting strategy achieves a better balance between visual fidelity and robustness.

B.3 Evaluation Metrics and Protocols

We employ three quantitative metrics that capture different aspects of avatar editing quality, complemented by a user study for perceptual evaluation. All metrics are computed on per-view edited images: for VTON360 we use the raw network outputs, for AvatarMix we use the images refined by SeamFix and FullbodyFix, and for TIP-Editor we use the edited rendered images produced by their pipeline.

DINO Similarity for Editing Target. To assess how well each method preserves the appearance of the region it edits, we compute DINO [44] feature similarity, akin to the garment similarity metric used in VTON360. However, since the methods we compare have different editing targets (head and body for our method, upper-body garment for VTON360, and head only for TIP-Editor), we corre-

spondingly adjust the target for computing DINO similarity for fair comparison as follows. For VTON360, whose target is the upper garment only, we compare its edited images against the corresponding front and back garment references using a garment-only segmentation mask. Because VTON360’s try-on results may exhibit substantial pose changes relative to the garment images, we restrict this comparison to front (0°) and back (180°) views, which are the most geometrically aligned and thus conservative in favor of VTON360. For AvatarMix, whose target is the fully clothed body, we compare 36 edited views against renders of the ground-truth model avatar using a clothed-body mask, making the evaluation stricter despite the smaller geometric changes introduced by body reshaping. TIP-Editor performs head-only replacement and leaves the garment and body unchanged in our setting, so this metric is not applicable to TIP-Editor.

Head and Neck DINO Similarity. To evaluate facial identity preservation and the seamlessness of the neck region, we compute DINO feature similarity on a head-and-neck segmentation mask between the edited images and the ground-truth user avatar. We evaluate this metric over 36 views for all three methods. This protocol is disadvantageous to AvatarMix: after editing, the head and body align with the model’s pose, so self-occlusions can differ between the edited and user avatar, which tends to reduce similarity scores even when identity is preserved. Despite this, AvatarMix still achieves the highest head-and-neck DINO similarity.

Warping-based RMSE. To quantify multi-view consistency, we use a warping-and-RMSE metric computed directly on the edited images instead of the CLIP Direction Consistency Score [24] used in VTON360. Directional CLIP evaluates whether appearance changes between neighboring views are similar before and after editing, which is suitable when edits mainly affect texture while pose is fixed, as in VTON360’s original setting. In our case, both garment appearance and the user’s pose can change after editing due to our compositional identity transfer approach; we observe disagreement between Directional CLIP and human judgments of consistency in this setting. We therefore adopt a more direct image-space measure: following the public implementation from the work of Asim *et al.* [2], we first estimate dense 2D correspondences between neighboring views and then measure the root-mean-squared error between one view and the other warped into its coordinate frame. Lower values indicate that details and geometry are stable across viewpoints. We report this metric for all three methods using their respective edited images across all viewpoints.

C. Limitations and Future Work

C.1 Limitations

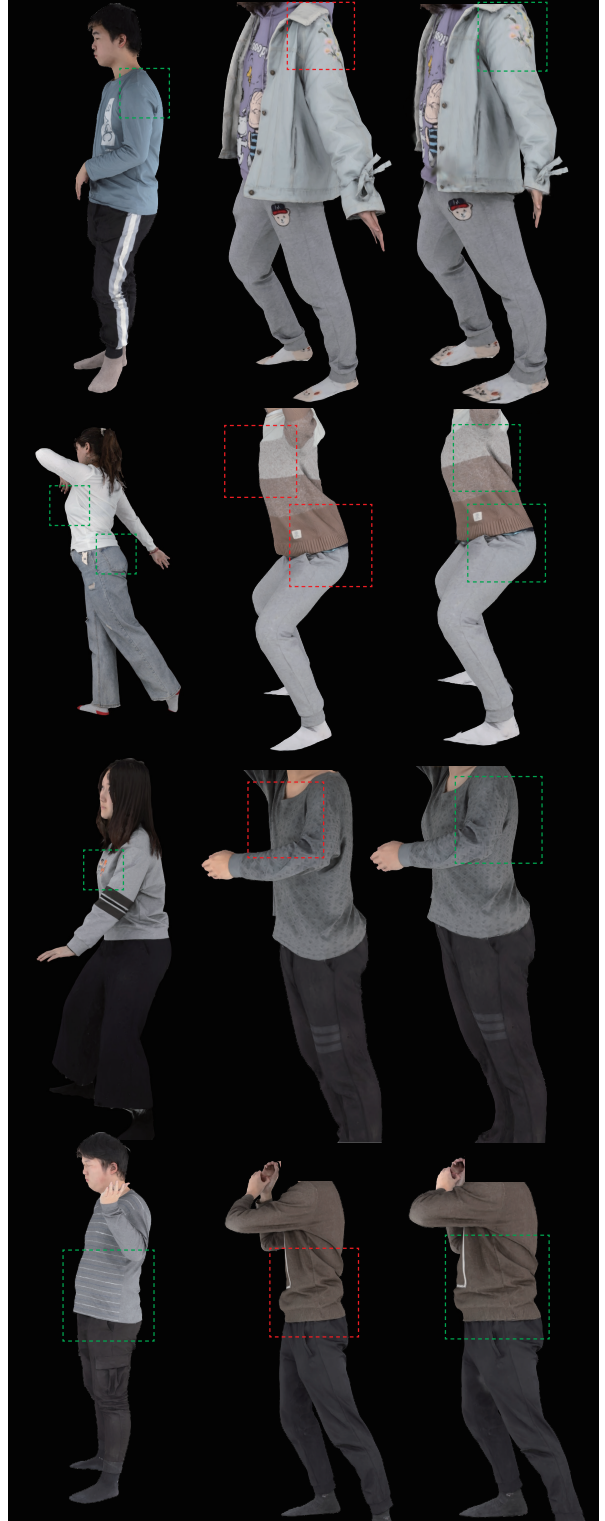
While AvatarMix achieves strong results on THuman2.0, several limitations remain. First, our current GSReshape design does not explicitly model detailed hand shape adaptation, which can lead to mismatches at extreme body shape differences. In addition, very loose garments or highly complex accessories may challenge the underlying garment re-targeting, occasionally producing wrinkles or folds that differ from the original model.

C.2 Future Work

One of the future works is exploring more diverse datasets beyond THuman2.0 to assess generalization across broader clothing styles. Another promising direction is to explore avatar reposing from reconstructed 3D Gaussians (existing reposing works often take a monocular video [21] or multi-view videos [36, 38] as input), extending user’s control on avatar pose after garment personalization.

D. Additional Qualitative Results

We provide additional multi-view comparisons on THuman2.0 not shown in the main paper (Fig. 9), demonstrating that AvatarMix maintains facial identity, seamless neck transition, and garment fidelity across challenging poses and lighting conditions compared to VTON360 and TIP-Editor. We also show additional ablation results for GSReshape (Fig. 8), illustrating how our retargeting module successfully adapts garments from the model avatar to the user’s body shape while preserving garment details. We provide the 360-degree videos of ours and comparison methods, rendered with updated (for ours and TIP-Editor) or reconstructed (for VTON360) Gaussian avatars in the supplementary material.



Body shape reference w/o GSReshape w/ GSReshape

Figure 8. Additional ablation on GSReshape. We visualize the effect of our body reshaping module by comparing the model avatars without GSReshape versus with GSReshape. As shown in the with GSReshape results, the garment adapts smoothly to the user’s body shape while preserving details after the body reshaping.

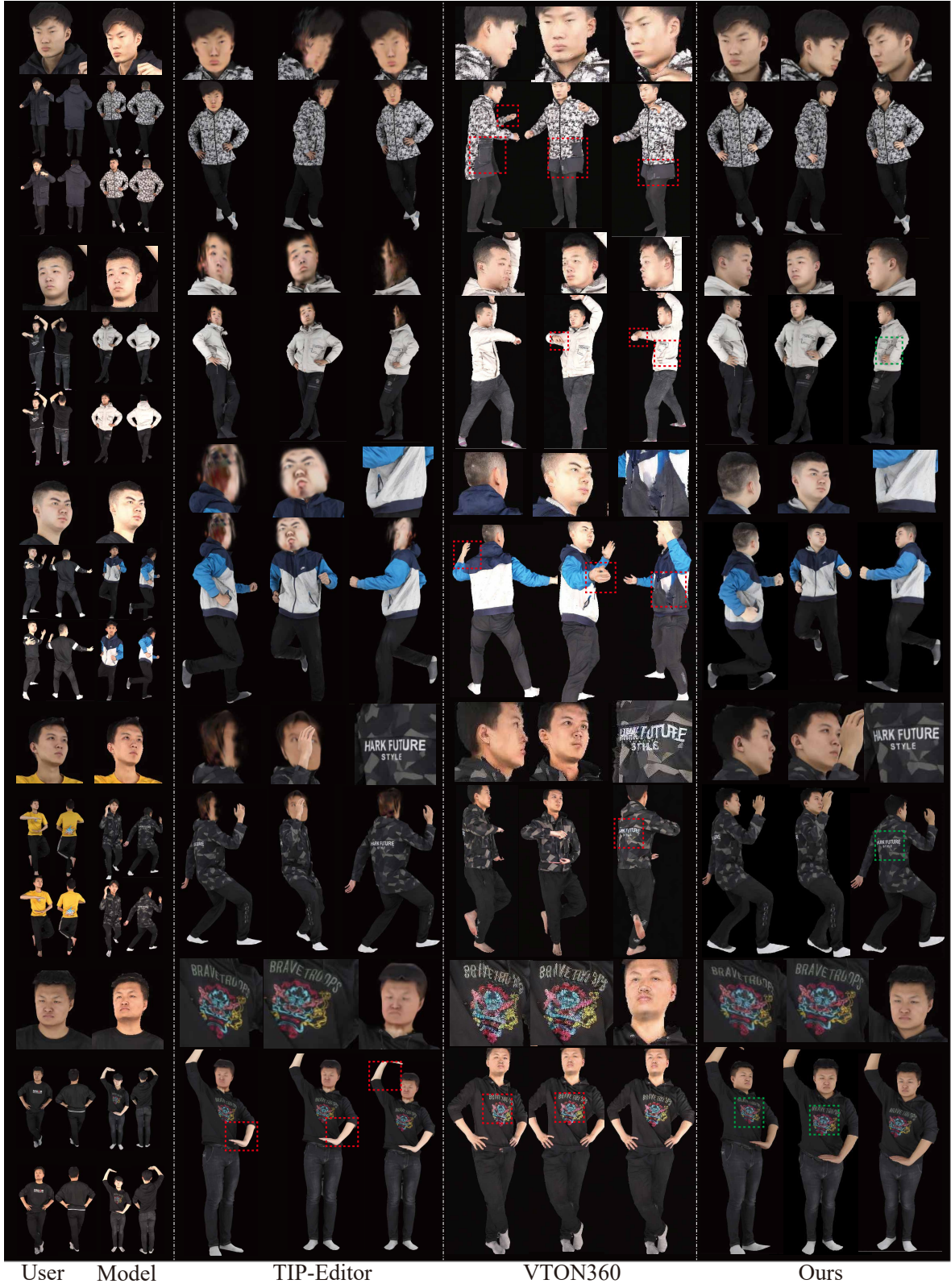


Figure 9. **Additional comparisons with THuman2.0.** We compare AvatarMix with baselines on more user-model pairs, demonstrating superior preservation of identity and outfit across diverse views.