CLOSELY INTERACTING 3D CLOTHED HUMAN GENERATION WITH A TIGHTNESS SIGNAL

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce *interacting clothed human generation*, generating a pair of 3D clothed humans in close contact such as hugging, holding hands, or dancing, from text and a target interaction pose. Existing work either generates interacting bodies without clothing or reconstructs clothed people in contact from a real image, while single clothed-human generators produce an isolated person with limited text control and fidelity. No prior method generates novel *interacting clothed* humans. We propose a training-free pipeline that generates each clothed person in 2D, so a text prompt controls diverse subjects and outfits at high fidelity. It then reaches the interaction pose in 3D for exact contact, and resolves the garment interpenetration. The tightness signal drives both the reposing and the penetration solution through one confidence-weighted variational solver. First, *tightness-aware skinning* uses the signal's direction to recover cloth-to-body correspondences and its magnitude as a per-vertex confidence in the variational solver for the skinning weights, enabling template-free reposing without the loose-garment artifacts of naive weight transfer. Second, *penetration-aware combination* solves for a smooth per-vertex displacement field that pushes interpenetrating garments apart without distorting them, with the tightness signal making the looser garment yield more. In the experiments on diverse datasets, our reposing lowers cloth distortion on every measure, reducing out-of-plane bending by $5.6\times$ over naive skinning, and our combination removes 94% of the interpenetration volume without distorting the garments. As a complete pipeline, it generates diverse, text-driven pairs of interacting clothed humans whose individual people surpass state-of-the-art single-person generators in appearance and geometry.

1 INTRODUCTION

We address *interacting clothed human generation*, generating a pair of 3D clothed humans in close contact such as hugging, holding hands, or dancing, from text and a target interaction pose (Fig. 1). Such contact-rich configurations are the building blocks of populated 3D human scenes for games and film, and can be animated and rendered into video. Yet generating two clothed people in interaction remains far less explored than the single-person case and is non-trivial, as it must jointly handle the interaction pose, the inter-person contact, and the clothing layered on both bodies.

Existing methods address only parts of this problem. Interaction *generation* samples novel interacting bodies from a generative prior over the parametric body, so it yields pose and shape but no clothing (Müller et al., 2024; Liu et al., 2025b). Clothed multi-person *reconstruction* (Kim et al., 2026) instead recovers garment-level geometry for multiple people in contact from a single real image, and so reproduces an observed moment rather than creating novel people or outfits. Single clothed-human generators, by contrast, produce an isolated person with limited text control and fidelity (Zhang et al., 2024a; 2025; Abdal et al., 2024; Zhang et al., 2023b; Liu et al., 2024). None of these approaches generates novel *interacting clothed* humans.

We propose a training-free pipeline that independently generates each clothed person from text and assembles the two into the target interaction (Fig. 2). Our pipeline rests on two design choices. First, we generate each person in 2D and lift it to 3D, so that a text prompt controls diverse, high-fidelity subjects that native 3D generation cannot. Second, we reach the interaction pose by reposing in 3D rather than posing the image, so the two people meet in the exact contact that a close interaction

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.

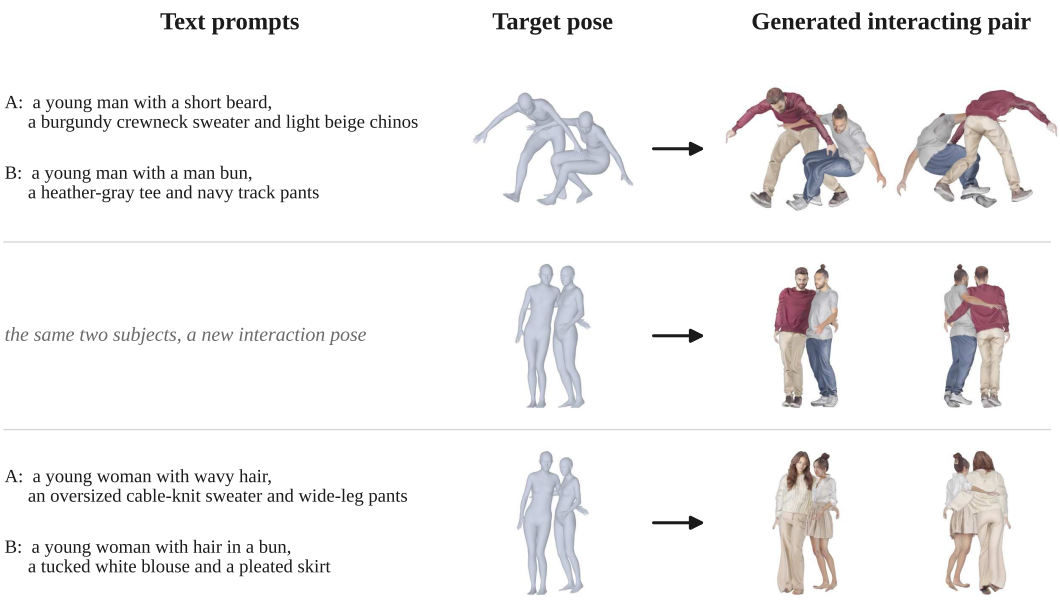


Figure 1: **Interacting clothed human generation.** From two text prompts and a target interaction pose (gray bodies, sampled from FlickrCI3D), our pipeline generates a pair of 3D clothed humans in close contact, shown front and back. Rows 1–2 show the same subjects in two interaction poses (pose control), and rows 2–3 share one interaction pose, varying the prompts (text control).

requires (Sec. 3.1). We instantiate these with an off-the-shelf text-to-image model (Black Forest Labs, 2024), a single-view lifter (Chen et al., 2025), and ETCH (Li et al., 2025a) for a per-vertex tightness vector. Two geometric steps then form the interaction, both driven by the same tightness signal through one confidence-weighted variational solver. *Tightness-aware skinning* reposes each template-free mesh to the target pose, using the signal's direction to recover the cloth-to-body correspondence and its magnitude as a per-vertex confidence in the solver for the skinning weights. This avoids the *webbing* and *splitting* artifacts of naive weight transfer on loose garments. *Penetration-aware combination* then places the two in a shared frame and resolves their interpenetration by solving for a smooth per-vertex displacement field, so the garments separate without distortion. The tightness signal sets the interface so the looser garment yields while the tighter one keeps its fit. In the experiments on 4D-DRESS (Wang et al., 2024b) and FlickrCI3D (Fieraru et al., 2020), our reposing lowers cloth distortion on every measure, reducing out-of-plane bending by $5.6\times$ over naive skinning, and our combination removes 94% of the interpenetration volume without distorting the garments. As a complete pipeline, it generates diverse, text-driven pairs of interacting clothed humans whose individual people surpass state-of-the-art single-person generators in appearance and geometry (Sec. 4.2, App. D). Our contributions are summarized as follows:

- **Interacting clothed human generation:** the first pipeline, generating each person in 2D for text control over diverse subjects and outfits at high fidelity, then lifting to 3D and reaching the interaction pose by reposing for exact contact.
- **Tightness-aware skinning:** a template-free reposing that uses a tightness signal's direction to recover the cloth-to-body correspondence and its magnitude as a per-vertex confidence in the solver, removing the webbing and splitting artifacts of naive weight transfer.
- **Penetration-aware combination:** a method that places the two people in a shared frame and resolves their cloth interpenetration by solving for a smooth displacement field, so the garments separate without tearing, with the tightness signal deciding how much each garment yields.

2 Related Work

Interacting human generation and reconstruction. Prior work on more than one person in close interaction has dealt mostly with the bare body shapes up to SMPL (Loper et al., 2015). Many methods learn generative priors over two-person interaction, sampling either a static contact pose

or a motion sequence from noise, text, or a single pose (Müller et al., 2024; Liang et al., 2024; Liu et al., 2025b; Shafir et al., 2024; Javed et al., 2025; Ponce et al., 2024), but none produce clothing. A separate line reconstructs clothed people in contact from a real image, video, or multi-view capture (Mustafa et al., 2021; Cha et al., 2024; Zheng et al., 2021a; Jiang et al., 2026; 2024; Kim et al., 2026), reasoning about occlusion, depth ordering, and physical contact to separate the subjects. These methods recover observed subjects faithfully but cannot synthesize new ones, leaving the generation of novel interacting clothed humans unaddressed.

Single clothed-human generation and reconstruction. Single-image reconstruction recovers a clothed mesh through pixel-aligned or parametric-conditioned implicit functions, normal integration, or text guidance (Saito et al., 2019; 2020; Zheng et al., 2021b; Xiu et al., 2022; 2023; Zhang et al., 2024b; Huang et al., 2024). More recent methods use image-conditioned diffusion priors to hallucinate the unseen regions (Ho et al., 2024; Li et al., 2025b). We build on SyncHuman (Chen et al., 2025), which synchronizes a 2D multi-view generator with a 3D-native model for single-image reconstruction. In parallel, text-driven and unconditional generators synthesize clothed avatars on NeRF, mesh, or Gaussian representations (Hong et al., 2022; Kolotouros et al., 2023; Liu et al., 2024; Kim et al., 2023; Liao et al., 2024; Dong et al., 2024; Wang et al., 2024a; Chen et al., 2022; Zhang et al., 2023b; Hu et al., 2024; Zhang et al., 2024a; 2025; Abdal et al., 2024). All of these produce a single isolated person, with limited text control and fidelity (Sec. 4.2). We instead compose two such single-person generations into one interaction, reposing each to the target pose and resolving the interpenetration that single-human generation never faces.

Skinning and reposing of clothed humans. Given the SMPL skeleton, reposing a clothed mesh reduces to assigning per-vertex skinning weights, classically computed from mesh geometry relative to an embedded skeleton (Jacobson et al., 2011; Dodik et al., 2024; Baran & Popović, 2007; Dionne & de Lasa, 2013) or predicted by learned neural rigging, including template-free variants (Xu et al., 2020; Liu et al., 2025a; Guo et al., 2025), though these rig a static asset rather than reposing it onto a given body pose. Weights can instead be transferred from the fitted body by nearest-point matching and smooth inpainting, as in RSWT (Abdrashitov et al., 2023), but this correspondence fails on loose clothing. Learned draping instead predicts garment deformation and skinning from data (Jiang et al., 2020; Bertiche et al., 2021a;b; Patel et al., 2020; De Luigi et al., 2023), with GAPS (Chen et al., 2024) using a body-garment proximity measure closest to our tightness cue, but each is garment-specific, and several require a garment template. Learned avatar methods instead learn a canonical shape and skinning from scanned surfaces (Saito et al., 2021; Chen et al., 2021), but require per-subject optimization, and free-form generators target loose clothing (Ye et al., 2025). We instead use the tightness signal from ETCH to repose a single fused mesh with no template and no training. Cloth simulation could add physical wrinkles (Li et al., 2020; Grigorev et al., 2023; Santesteban et al., 2022) but needs a separated garment and material parameters our generated meshes lack, so we deform the surface kinematically and leave physical relaxation to future work.

Interpenetration and collision resolution. Resolving interpenetration between surfaces is a long-standing problem. Physical simulation prevents or resolves collisions, from IPC (Li et al., 2020; 2021), which guarantees intersection-free contact, to learned neural simulators (Grigorev et al., 2024) that recover from intersections after they occur, all at the cost of a garment template, a body collider, and simulation state. In 3D human fitting, interpenetration is instead discouraged by a soft penalty on the parametric body (Bogo et al., 2016; Pavlakos et al., 2019; Hassan et al., 2019), not the garment. Contour-minimization and global-analysis untanglers (Volino & Magnenat-Thalmann, 2006; Baraff et al., 2003) resolve intersections geometrically, training- and template-free like ours, but run as collision responses inside a cloth simulation. Layered-clothing methods (Buffet et al., 2019; Aggarwal et al., 2022; Lee et al., 2023) resolve penetration between garment layers, but require a garment template, an explicit layer order, or training. We instead reconcile two independently generated humans in a single static solver, with no simulation, collider, template, or layer input.

3 METHOD

Our pipeline turns text and a target interaction pose into a pair of clothed 3D humans (Fig. 2). Each person is first generated and lifted to 3D by off-the-shelf models (FLUX, SyncHuman, ETCH), producing a single fused clothed mesh with a per-vertex tightness vector. Two geometric steps then assemble the interaction. *Tightness-aware skinning* (Sec. 3.3) reposes each person to the interaction

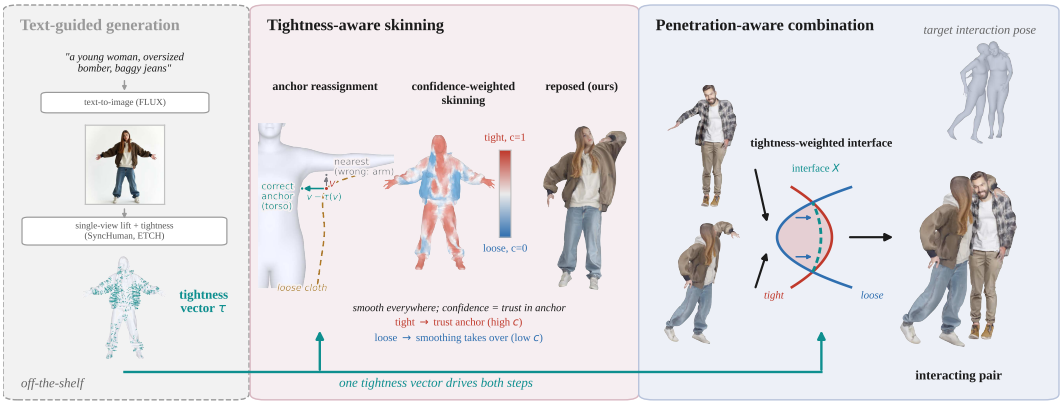


Figure 2: **Pipeline overview.** From one garment prompt per person and a target interaction pose, off-the-shelf models generate each person as a canonical clothed 3D mesh carrying a per-vertex tightness vector τ (FLUX for text-to-image, SyncHuman for single-view lifting, ETCH for tightness). Two geometric steps then assemble the interaction: *tightness-aware skinning* reposes each person to the target pose by anchor reassignment and a confidence-weighted skinning solve, and *penetration-aware combination* places the two in a shared frame and resolves their cloth interpenetration across a tightness-weighted interface X. The same tightness signal drives both steps.

pose, and *penetration-aware combination* (Sec. 3.4) places the two in a shared frame and resolves their interpenetration. Both are driven by the same tightness signal through one variational solver.

3.1 Design Rationale

Generating each person in 2D. We synthesize each person with a text-to-image model and lift the image to 3D, rather than generating natively in 3D, which gives two things a 3D generator cannot. **First, fidelity.** An image model is trained on far more data than any 3D generator, so lifting carries high-fidelity appearance and geometry into the mesh, whereas 3D-native generators built on scarce 3D data are markedly lower-fidelity. **Second, text control.** The prompt directs each subject and garment, so one pipeline produces people across ages, body types, and outfits from text alone, while feed-forward 3D generators lack text control and score-distillation text-to-3D is slow and low-fidelity. Lifting a 2D generation gives both fidelity and text control, which native 3D generation does not (Fig. 9). We reconstruct each person in a canonical, fully-visible pose, where the lift is most faithful and which is exactly what the reposing stage needs.

Reaching the interaction pose in 3D. A close interaction is defined by where the two clothed surfaces meet, so the target pose and contact must be reached exactly. Reposing the canonical asset in 3D places both people at the exact interaction pose in one shared frame, so their relative placement follows the target interaction directly. By contrast, applying the pose in the image plane does not reach this control. A pose-conditioned image model struggles to produce a person in an uncommon interaction pose (FLUX with ControlNet (Zhang et al., 2023a)). A 2D reposer can reach the pose instead (ReImagine (Sun et al., 2026)), but its posed image self-occludes, so lifting it degrades the hidden regions and leaves the asset without accurate geometry (Fig. 7, detailed in App. A).

3.2 Preliminaries

Notation. We write v for the vertices of the fused clothed mesh, indexed by i. Each body point b on the fitted SMPL model carries a skinning weight $W_{\text{smpl}}(b) \in \mathbb{R}^{24}$, one per joint, and reposing applies the per-joint rigid transforms that map the source pose A_j^{cur} to the target pose A_j^{new}. At each vertex, ETCH provides a tightness vector $\tau(v)$ from the cloth to the underlying body, so that $v + \tau(v)$ is the corresponding body point and $\|\tau(v)\|$ measures looseness.

Naive LBS. Cloth has no skinning weights, so each vertex borrows from a body point b, $W(v) = W_{\text{smpl}}(b)$, and is deformed by linear blend skinning (LBS):

$$v' = \sum_j W_j(v) \left(A_j^{\text{new}} (A_j^{\text{cur}})^{-1}\right) v. \tag{1}$$

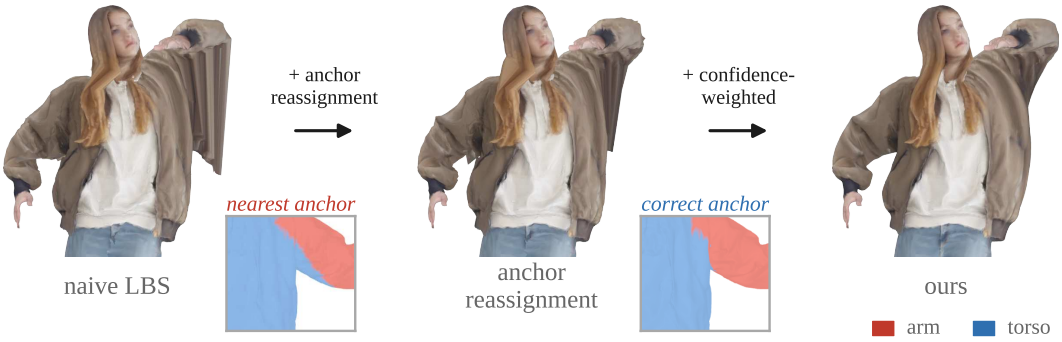


Figure 3: **Tightness-aware reposing.** A loose jacket reposed to an interaction pose. Naive LBS drags the armpit panel into a *webbing* artifact. Anchor reassignment and the confidence-weighted solve fix it (*ours*). The insets show which joint the free-hanging armpit cloth borrows weight from. Naive assignment picks the nearer arm, while reassignment picks the torso it actually hangs from.

The naive choice is the nearest body point, $b_{\text{dist}} = \arg\min_b \|v - b\|$, whose failures on loose garments motivate our tightness-aware skinning (Sec. 3.3).

Smoothness energy. To fill in and smooth the skinning weights, we use a standard smoothness operator Q (Bang & Lee, 2018), built on the cotangent Laplacian L and a mass matrix M (Pinkall & Polthier, 1993), and used by RSWT (Abdrashitov et al., 2023) for weight inpainting,

$$Q = \underbrace{L}_{\text{Dirichlet}} + \underbrace{L\,M^{-1}L}_{\text{bi-Laplacian}}, \qquad E_{\text{smooth}}(W) = W^{\top} Q\, W. \tag{2}$$

$E_{\text{smooth}}(W)$ measures how non-smooth a per-vertex field W is, so minimizing it fills the field smoothly. The Dirichlet term interpolates and the bi-Laplacian term keeps the result curvature-continuous. The full construction and the weight-inpainting solve are in App. B.

3.3 Tightness-Aware Skinning

Nearest-point transfer fails on loose garments in two ways long observed for body-to-cloth weight transfer. When loose cloth bridges two body parts, such as an armpit panel, nearest-point assignment binds it to the nearer one and drags it along the wrong joint into a flat sheet across the gap, the *webbing* artifact (Abdrashitov et al., 2023) (Fig. 3). A skirt spans both legs as a single tube, so each side is pulled toward its nearer leg and the skirt tears open between the legs as they separate, the *splitting* artifact (Chen et al., 2024; Bertiche et al., 2021b; Jiang et al., 2020). Our tightness-aware skinning fixes both, using the tightness signal in two roles: its *direction* reassigns the skinning anchor (correspondence), and its *magnitude* sets a per-vertex confidence that weights this anchor against the smoothness energy Q in a variational solver (continuity). We detail the two mechanisms next.

Anchor reassignment. The nearest body point sends free-hanging cloth to the wrong joint. We instead borrow the weight from the corresponding body point, $v+\tau(v)$, giving the reassigned anchor $W^0(v) = W_{\text{smpl}}(b_{\text{tip}})$ where

$$b_{\text{tip}} = \arg\min_b \|(v + \tau(v)) - b\|\,. \tag{3}$$

This recovers the cloth-to-body correspondence distance alone misses, fixing the webbing artifact.

Confidence-weighted skinning. RSWT freezes the matched, tight vertices as a hard boundary (App. B.1), so its smoother never regularizes them and the lumpiness they induce under naive transfer stays frozen in place. Meanwhile the tip anchor becomes unreliable as cloth moves away from the body. We therefore make *every* vertex a variable, softly pulled toward its anchor W^0 in proportion to a per-vertex confidence $c(v)$, while the *same* smoothness energy Q of Eq. (2), weighted by λ, regularizes the whole field:

$$W = \arg\min_W \sum_i c(v_i)\ \left\|W_i - W^0(v_i)\right\|^2 \ +\ \lambda\, W^{\top} Q\, W. \tag{4}$$

This is a sparse, symmetric positive-definite system whose solution, with $C = \operatorname{diag}(c)$, satisfies

$$(C + \lambda Q)\, W = C\, W^0. \tag{5}$$

Because Q now acts on every vertex, the confidence controls a smooth transition rather than the hard cutoff. Tight vertices stay close to the transferred detail, moderately loose vertices are interpolated, and at the loosest vertices ($c \to 0$) the data term vanishes, so the solve reduces to a pure Q-smoothing inpaint. The smoother Q itself is unchanged from RSWT. The confidence is a raised-cosine gate on the looseness, $c(v) = \frac{1}{2}\big(1 + \cos(\pi\ \mathrm{clip}(\|\tau(v)\|\,/\tau_{\max}, 0, 1))\big)$, high where the cloth is tight and the tip anchor is reliable, and falling to zero past the cutoff $\tau_{\max}$ as the cloth loosens (App. B.2).

Although each anchor is a single body point, the solve blends it with the anchors of neighboring cloth vertices, an implicit multi-point average on the *cloth surface*. Averaging over surface neighbors keeps our skinning correspondence-aware, rather than mixing body points that are merely close in space as a body-space kernel such as GAPS would. The operator derivation is in App. B.3.

3.4 Penetration-Aware Combination

Two reposed people are placed in a shared frame following the target interaction, which can leave their bodies overlapping and their garments interpenetrating. After separating the bodies with a minimal pose correction (App. C.1), we remove the garment interpenetration with the confidence-weighted solve of Eq. (5), now acting on per-vertex displacements d instead of skinning weights.

Tightness-weighted interface. Let A and B be the two garments. We resolve their interpenetration against an explicit interface placed between them, and touch only the overlap region $\mathcal{O}$, which we identify with the generalized winding number (Jacobson et al., 2013). Writing $\phi_A(p)$ and $\phi_B(p)$ for the distance from a point p to garment A and B, we place the interface X within $\mathcal{O}$ as the level set where these two distances balance in proportion to how much each garment should yield,

$$X = \big\{\, p \in \mathcal{O} :\ (1 - w(p))\,\phi_A(p) = w(p)\,\phi_B(p) \,\big\}, \quad w(p) = \frac{\|\tau_A(p)\|}{\|\tau_A(p)\| + \|\tau_B(p)\|}, \tag{6}$$

with $\|\tau_A(p)\|, \|\tau_B(p)\|$ the local looseness of the two garments. The looser garment takes the larger weight and travels further, so X sits closer to the tighter garment, which stays put (Fig. 2). Setting $w = \frac{1}{2}$ recovers the plain equidistant surface, which we ablate in App. C.2.

Displacement solve. For each penetrating vertex we set a target $d^0(v)$, the displacement that moves it just past the interface X, its length equal to the penetration depth. A direct push that moved each vertex onto its target would tear the cloth at the patch boundary. We instead solve for a smooth displacement field d with the same system as reposing (Eq. (5)), on displacements instead of weights,

$$(C + \lambda Q)\, d = C\, d^0. \tag{7}$$

Penetrating vertices are assigned a high confidence and pushed onto their targets, while the surrounding cloth is regularized toward zero displacement under the tightness confidence and feathered to zero at the patch boundary so it blends without a crease. Because the solve smooths the displacement, vertices approach their targets rather than reaching them exactly, leaving a residual overlap that we reduce by iterating the solve and keeping the least-residual pass (App. C.3).

4 Experiments

4.1 Experimental Setup

Datasets. Reposing uses 4D-DRESS (Wang et al., 2024b), taking 15 outfits binned by measured looseness into loose, medium, and tight (five each). For interpenetration, these outfits are placed in two-person interaction poses sampled at random from FlickrCI3D (Fieraru et al., 2020).

Baselines. With no baseline for the task, we compare against single-human 3D generators, applied to each person independently: the unconditional GETAvatar (Zhang et al., 2023b), E3Gen (Zhang et al., 2024a), LayerAvatar (Zhang et al., 2025), and GSM (Abdal et al., 2024), evaluated quantitatively, together with the text-conditioned HumanGaussian (Liu et al., 2024) shown qualitatively. For reposing, we compare our tightness-aware skinning against naive weight transfer, RSWT (Abdrashitov et al., 2023), and GAPS (Chen et al., 2024).

Metrics. For generation, which has no paired ground truth, we score every method against real captures, using photographs (single-person SHHQ (Fu et al., 2022), two-person FlickrCI3D) for

Table 1: **Single-human generation.** FID and KID for appearance (RGB renders, reliable against the photographs) and geometry (normal-map renders, subscript *norm*, reliable against the scans), both reported on SHHQ photographs and 4D-DRESS scans. N=5,000 for SHHQ and N=9,216 for 4D-DRESS (64 identities $\times$ 8 poses $\times$ 18 views). KID $\times 10^{-3}$, lower is better, **best** and second.

	SHHQ				4D-DRESS			
Method	FID↓	KID↓	FID_{norm} ↓	KID_{norm} ↓	FID↓	KID↓	FID_{norm} ↓	KID_{norm} ↓
GETAvatar	113.0	127.5	72.5	68.1	64.2	54.8	38.1	36.7
GSM	53.2	45.9	–	–	77.3	66.2	–	–
E3Gen	71.9	67.6	51.0	47.9	47.1	34.2	26.5	24.6
LayerAvatar	59.2	52.4	49.0	47.3	44.7	34.0	34.5	34.4
Ours	**38.0**	**32.3**	**28.2**	**22.8**	**44.1**	**30.3**	**16.1**	**14.3**

Table 2: **Two-person generation.** FID and KID for appearance and geometry (subscript *norm*), computed as in Table 1, against FlickrCI3D photographs and Hi4D scans. FlickrCI3D is reported with KID only, being too small for a reliable FID. N=1,281 for FlickrCI3D and N=7,200 for Hi4D (100 sequences $\times$ 4 poses $\times$ 18 views). KID $\times 10^{-3}$, lower is better, **best** and second.

	FlickrCI3D		Hi4D			
Method	KID↓	KID_{norm} ↓	FID↓	KID↓	FID_{norm} ↓	KID_{norm} ↓
GETAvatar	186.4	72.4	96.6	120.9	48.0	55.5
GSM	42.5	–	101.0	106.4	–	–
E3Gen	60.7	54.8	77.4	83.8	42.5	49.7
LayerAvatar	74.4	**40.2**	73.0	72.3	**29.7**	**31.7**
Ours	**35.5**	45.5	**50.8**	**39.4**	31.2	34.4

appearance and 3D scans (single-person 4D-DRESS, two-person Hi4D (Yin et al., 2023)) for geometry, from front-view RGB renders and 18-azimuth normal-map renders respectively. On the photo datasets, which lack ground-truth normals, we additionally report a front-facing pseudo-normal metric (App. D.2). We detail this protocol in App. D.1. Reposing distortion follows prior garment-draping evaluations (Chen et al., 2024; De Luigi et al., 2023) and scores the reposed surface intrinsically, with no ground-truth posed scan. The mean edge-strain ε_e, the mean area-strain ε_a, the 99.5th-percentile stretch (worst-case edge elongation), and the bending deviation E_{bend} (change in dihedral angle) each measure how far the reposing stretches or bends the cloth relative to its source. We measure interpenetration as the intersection volume between the two people's garments, computed with generalized winding numbers.

Implementation. The pipeline is training-free with no per-subject optimization and assembles off-the-shelf models. On a single RTX 3090, generation and reconstruction take about four minutes per person, paid only once, while each interacting pair adds only about 25 s for our reposing and penetration solution.

4.2 Generation Quality

Generating each person in 2D. Against recent single-human 3D generators, our method attains the lowest FID and KID on every metric (Table 1), leading on appearance (SHHQ FID 38.0 against 53.2) and on geometry (4D-DRESS normal-FID 16.1 against 26.5), with the largest margin on geometry. These baselines are all unconditional. Text-driven 3D generation is too slow to sample at FID scale, so we compare it qualitatively with HumanGaussian (Fig. 9). It produces cartoonish, lower-fidelity results, while our route keeps both prompt control and sharp geometry.

Two-person generation. No prior method generates interacting clothed humans, so we apply the same single-human generators to the two-person setting. Our pipeline produces a wide range of interacting pairs across subjects, outfits, and poses (Fig. 4). Quantitatively, these pairs have the lowest appearance KID on FlickrCI3D photographs, while their geometry is close to the best baseline on Hi4D scans (Table 2). Because both people are placed in a matched pose, these distances gauge the quality of each posed person within the interaction rather than the interaction itself. The quality specific to an interaction, its residual cloth penetration, is evaluated in Sec. 4.4.



Figure 4: **Two-person generation gallery.** Interacting clothed pairs generated by our pipeline, spanning diverse ages, body types, and outfits (including loose garments) across varied interaction poses, each in plausible close contact. Front view. More samples in Fig. 10.

Table 3: **Cloth reposing quality** on identical fused meshes, 15 outfits at 8 target poses each. The four distortion measures are defined in Sec. 4.1, with strains in percent and E_{bend} in degrees, lower better. The right block reports per-bin edge-strain ε_e by measured looseness, five outfits per bin.

Method	$\varepsilon_e \downarrow$	stretch$\downarrow$	$\varepsilon_a \downarrow$	$E_{\text{bend}} \downarrow$	loose	med	tight
naive	13.4	3.4	22.8	7.7	16.4	13.4	10.6
RSWT	10.5	2.5	16.4	5.1	11.0	10.5	10.1
GAPS	9.6	2.6	14.4	2.4	11.3	9.8	7.7
Ours	**8.2**	**1.8**	**12.0**	**1.4**	**9.3**	**8.3**	**6.9**

4.3 Cloth Reposing Quality

We evaluate our tightness-aware skinning in isolation on 4D-DRESS. For each outfit we take the frame closest to a canonical pose as the source, a single scan with no template or skinning weights, exactly our method's input. We transfer weights onto it under four fields (naive LBS, RSWT, GAPS, and ours), so that only the weight field differs (Table 3).

The methods fall in a clean, monotone order, naive $>$ RSWT $>$ GAPS $>$ ours, with ours lowest in every metric and every looseness bin. The gap is widest in bending, an out-of-plane curvature the strain terms cannot capture, where ours is $5.6\times$ smaller than naive LBS, as our smoother removes the spurious creasing that nearest-point weights inject. The ranking and this margin also hold at full scale, on all 64 clean-source outfits at 32 target poses each ($2{,}048$ cases, a seventeen-fold increase, Table 4). The correspondence gain from anchor reassignment is stretch-neutral and is instead shown qualitatively on the webbing case (Fig. 3). Our soft confidence further removes the facets RSWT's hard freeze locks into loose cloth, while preserving tight detail (Fig. 5).

4.4 Penetration Solution

Since no method generates interacting clothed humans, there is no external baseline, and existing penetration resolvers need a garment template or simulation state that our generated meshes lack, so we compare the interpenetration with the resolver off (the raw assembly) and on, following Hi4D's intersection-volume metric (Yin et al., 2023). Interaction poses are drawn at random from Flickr-CI3D without curation, and each pair is reposed and placed by our pipeline.

On a fixed outfit across 150 random interaction poses, iterating the solve and keeping the least-residual pass removes $\mathbf{94.9}\%$ of the interpenetration volume ($165{,}450 \to 8{,}395\,\text{cm}^3$). Stratifying by tightness into six bins with a random outfit per case (900 cases, Table 5), the reduction is 92–96% in every bin (94.3% overall).

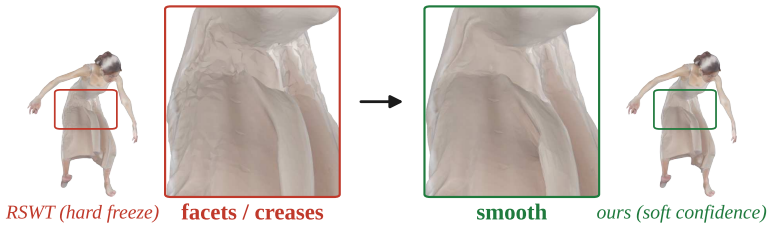


Figure 5: **Confidence-weighted skinning vs hard-boundary transfer.** A loose dress reposed with RSWT and our confidence-weighted solve. The boxed waist and upper-skirt region is enlarged at center. RSWT's hard freeze leaves faceted, creased cloth, whereas our soft per-vertex confidence lets the smoother regularize that frozen region, removing the facets while preserving tight detail.

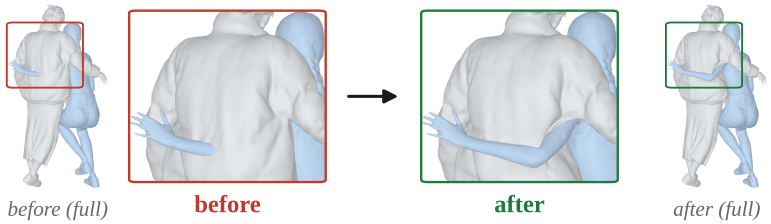


Figure 6: **Penetration-aware combination.** An interacting pair before and after our combination, with the boxed penetration region enlarged at center. Here an arm wraps over the other's back, pressing into their garment, and the solve separates them across the tightness-weighted interface, which sits closer to the tighter surface, so the looser garment absorbs the deformation smoothly.

Our solve keeps the garments essentially undistorted (Fig. 6). It carries the correction off the penetrating patch into the surrounding cloth, which absorbs the deformation smoothly rather than tearing at the boundary, so median edge-strain and bending stay near zero (App. C.3). A direct move of each penetrating vertex onto the interface, or dropping the boundary feathering, distorts the cloth far more, raising the worst-case edge stretch to 1.6–$2.1\times$ against $1.1\times$ for ours (Table 6, Fig. 8). Because the interface is tightness-weighted, the looser garment yields more, which we ablate against a plain equidistant split in App. C.2.

5 Conclusion

We introduced *interacting clothed human generation*, producing a pair of clothed 3D humans in close contact from text and a target pose. Our training-free pipeline generates each person in 2D under text control and reaches the interaction pose by 3D reposing, then assembles the two through two geometric steps, tightness-aware skinning and penetration-aware combination, both driven by the same tightness signal through one variational solver. Our reposing lowers cloth distortion on every measure and our combination removes almost all of the interpenetration without distorting the garments. As a complete pipeline, it generates diverse pairs of interacting clothed humans whose individual people surpass state-of-the-art single-person generators in appearance and geometry.

However, it does not model cloth physics. Because each garment is generated once and carried by skinning, its wrinkles move rigidly with the limbs rather than re-forming with the pose, and very loose cloth follows the body instead of draping freely. In the combination, a hand can still tunnel through the other person's garment or a small contact gap can open (App. E, Fig. 11). Still, these are cloth-level artifacts on an interaction whose pose and placement are otherwise accurate. Across diverse people and poses the pipeline reliably produces accurate interactions, offering a practical way to populate 3D scenes with clothed interacting people. Restoring pose-dependent wrinkling and draping, resolving the tunneling, and adding a contact-preserving term are the natural next steps.

AI use statement

In this work, we used generative AI tools, including large language models and an AI coding assistant, for three purposes, matching the categories declared in our submission form. First, to aid and

polish writing: editing and revising the text for clarity and concision and formatting bibliographic references. Second, for retrieval and discovery: searching for and identifying related work. Third, for research ideation and execution: assisting with implementation and code, and with the analysis and framing of our results and related-work comparisons. The core research idea, the method design, and the experimental protocol are the authors' own; all AI-assisted code was reviewed and tested, and every claim, number, figure, and citation was verified by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

Our method generates photorealistic 3D clothed humans. Like other generative human models, it could be misused to create synthetic imagery of people, for example for impersonation or misleading media; we encourage responsible use together with provenance, watermarking, and consent-based safeguards for any downstream deployment. Our generated subjects are produced from text prompts and are not real individuals. Our experiments use existing public research datasets (SHHQ, FlickrCI3D, 4D-DRESS, and Hi4D) under their respective licenses and terms, and involve no new human-subjects data collection. We have read and adhere to the ICLR Code of Ethics.

Reproducibility statement

The full construction of the tightness-aware skinning solve and the penetration-aware combination, including the energies, operators, confidence gate, and hyperparameters, is given in App. B and App. C. The evaluation protocol, covering datasets, metrics, pose-controlled FID/KID scoring, and the reposing and interpenetration measures, is specified in Sec. 4.1 and App. D.1. We will release our code and evaluation scripts to reproduce all reported results.

References

Rameen Abdal, Yifan Wang, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian Shell Maps for Efficient 3D Human Generation. In *CVPR*, 2024.

Rinat Abdrashitov, Kim Raichstat, Jared Monsen, and David Hill. Robust Skin Weights Transfer via Weight Inpainting. In *SIGGRAPH Asia Technical Communications*, 2023.

Alakh Aggarwal, Jikai Wang, Steven Hogue, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Layered-Garment Net: Generating Multiple Implicit Garment Layers from a Single Image. In *ACCV*, 2022.

Seungbae Bang and Sung-Hee Lee. Spline Interface for Intuitive Skinning Weight Editing. *ACM TOG*, 2018.

David Baraff, Andrew Witkin, and Michael Kass. Untangling Cloth. *ACM TOG*, 2003.

Ilya Baran and Jovan Popović. Automatic Rigging and Animation of 3D Characters. *ACM TOG*, 2007.

Hugo Bertiche, Meysam Madadi, and Sergio Escalera. PBNS: Physically Based Neural Simulator for Unsupervised Garment Pose Space Deformation. *ACM TOG*, 2021a.

Hugo Bertiche, Meysam Madadi, Emilio Tylson, and Sergio Escalera. DeePSD: Automatic Deep Skinning and Pose Space Deformation for 3D Garment Animation. In *ICCV*, 2021b.

Black Forest Labs. FLUX.1. `https://github.com/black-forest-labs/flux`, 2024.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *ECCV*, 2016.

Thomas Buffet, Damien Rohmer, Loïc Barthe, Laurence Boissieux, and Marie-Paule Cani. Implicit Untangling: A Robust Solution for Modeling Layered Clothing. *ACM TOG*, 2019.

Junuk Cha, Hansol Lee, Jaewon Kim, Nhat Nguyen Bao Truong, Jae Shin Yoon, and Seungryul Baek. 3D Reconstruction of Interacting Multi-Person in Clothing from a Single Image. In *WACV*, 2024.

Ruochen Chen, Liming Chen, and Shaifali Parashar. GAPS: Geometry-Aware, Physics-Based, Self-Supervised Neural Garment Draping. In *3DV*, 2024.

Wenyue Chen, Peng Li, Wangguandong Zheng, Chengfeng Zhao, Mengfei Li, Yaolong Zhu, Zhiyang Dou, Ronggang Wang, and Yuan Liu. SyncHuman: Synchronizing 2D and 3D Generative Models for Single-view Human Reconstruction. In *NeurIPS*, 2025.

Xu Chen, Yufeng Zheng, Michael J. Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable Forward Skinning for Animating Non-Rigid Neural Implicit Shapes. In *ICCV*, 2021.

Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J. Black, Andreas Geiger, and Otmar Hilliges. gDNA: Towards Generative Detailed Neural Avatars. In *CVPR*, 2022.

Luca De Luigi, Ren Li, Benoît Guillard, Mathieu Salzmann, and Pascal Fua. DrapeNet: Garment Generation and Self-Supervised Draping. In *CVPR*, 2023.

Olivier Dionne and Martin de Lasa. Geodesic Voxel Binding for Production Character Meshes. In *SCA*, 2013.

Ana Dodik, Vincent Sitzmann, Justin Solomon, and Oded Stein. Robust Biharmonic Skinning Using Geometric Fields. arXiv:2406.00238, 2024.

Junting Dong, Qi Fang, Zehuan Huang, Xudong Xu, Jingbo Wang, Sida Peng, and Bo Dai. TELA: Text to Layer-wise 3D Clothed Human Generation. In *ECCV*, 2024.

Zijian Dong, Xu Chen, Jinlong Yang, Michael J. Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to Generate 3D Avatars from 2D Image Collections. In *ICCV*, 2023.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-Dimensional Reconstruction of Human Interactions. In *CVPR*, 2020.

Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A Data-Centric Odyssey of Human Generation. In *ECCV*, 2022.

Artur Grigorev, Bernhard Thomaszewski, Michael J. Black, and Otmar Hilliges. HOOD: Hierarchical Graphs for Generalized Modelling of Clothing Dynamics. In *CVPR*, 2023.

Artur Grigorev, Giorgio Becherini, Michael J. Black, Otmar Hilliges, and Bernhard Thomaszewski. ContourCraft: Learning to Resolve Intersections in Neural Multi-Garment Simulations. In *SIGGRAPH Conference Papers*, 2024.

Zhiyang Guo, Jinxu Xiang, Kai Ma, Wengang Zhou, Houqiang Li, and Ran Zhang. Make-It-Animatable: An Efficient Framework for Authoring Animation-Ready 3D Characters. In *CVPR*, 2025.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D Human Pose Ambiguities with 3D Scene Constraints. In *ICCV*, 2019.

Hsuan-I Ho, Jie Song, and Otmar Hilliges. SiTH: Single-view Textured Human Reconstruction with Image-Conditioned Diffusion. In *CVPR*, 2024.

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *ACM TOG*, 2022.

Tao Hu, Fangzhou Hong, and Ziwei Liu. StructLDM: Structured Latent Diffusion for 3D Human Generation. In *ECCV*, 2024.

Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *3DV*, 2024.

Alec Jacobson, Ilya Baran, Jovan Popović, and Olga Sorkine-Hornung. Bounded Biharmonic Weights for Real-Time Deformation. *ACM TOG*, 2011.

Alec Jacobson, Ladislav Kavan, and Olga Sorkine-Hornung. Robust Inside-Outside Segmentation using Generalized Winding Numbers. *ACM TOG*, 2013.

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. InterMask: 3D Human Interaction Generation via Collaborative Masked Modeling. In *ICLR*, 2025.

Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. BCNet: Learning Body and Cloth Shape from a Single Image. In *ECCV*, 2020.

Yiming Jiang, Wenfeng Song, Shuai Li, and Aimin Hao. DECON: Reconstruction of Clothed-Geometric Multiple Humans from a Single Image via Geometry-Guided Decoupling. In *AAAI*, 2026.

Zeren Jiang, Chen Guo, Manuel Kaufmann, Tianjian Jiang, Julien Valentin, Otmar Hilliges, and Jie Song. MultiPly: Reconstruction of Multiple People from Monocular Video in the Wild. In *CVPR*, 2024.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Zhaoen Su, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for Human Vision Models. In *ECCV*, 2024.

Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. Chupa: Carving 3D Clothed Humans from Skinned Shape Priors using 2D Diffusion Probabilistic Models. In *ICCV*, 2023.

Gwanghyun Kim, Junghun James Kim, Suh Yoon Jeon, Jason Park, and Se Young Chun. Human Interaction-Aware 3D Reconstruction from a Single Image. In *CVPR*, 2026.

Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. DreamHuman: Animatable 3D Avatars from Text. In *NeurIPS*, 2023.

Dohae Lee, Hyun Kang, and In-Kwon Lee. ClothCombo: Modeling Inter-Cloth Interaction for Draping Multi-Layered Clothes. *ACM TOG*, 2023.

Boqian Li, Haiwen Feng, Zeyu Cai, Michael J. Black, and Yuliang Xiu. ETCH: Generalizing Body Fitting to Clothed Humans via Equivariant Tightness. In *ICCV*, 2025a.

Minchen Li, Zachary Ferguson, Teseo Schneider, Timothy Langlois, Denis Zorin, Daniele Panozzo, Chenfanfu Jiang, and Danny M. Kaufman. Incremental Potential Contact: Intersection- and Inversion-free, Large-Deformation Dynamics. *ACM TOG*, 2020.

Minchen Li, Danny M. Kaufman, and Chenfanfu Jiang. Codimensional Incremental Potential Contact. *ACM TOG*, 2021.

Peng Li et al. PSHuman: Photorealistic Single-image 3D Human Reconstruction using Cross-Scale Multiview Diffusion and Explicit Remeshing. In *CVPR*, 2025b.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-based Multi-human Motion Generation under Complex Interactions. *IJCV*, 2024.

Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! Text to Animatable Digital Avatars. In *3DV*, 2024.

Isabella Liu, Zhan Xu, Yifan Wang, Hao Tan, Zexiang Xu, Xiaolong Wang, Hao Su, and Zifan Shi. RigAnything: Template-Free Autoregressive Rigging for Diverse 3D Assets. *ACM TOG*, 2025a.

Shaowei Liu, Chuan Guo, Bing Zhou, and Jian Wang. Ponimator: Unfolding Interactive Pose for Versatile Human-human Interaction Animation. In *ICCV*, 2025b.

Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. HumanGaussian: Text-Driven 3D Human Generation with Gaussian Splatting. In *CVPR*, 2024.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *ACM TOG*, 2015.

Lea Müller, Vickie Ye, Georgios Pavlakos, Michael J. Black, and Angjoo Kanazawa. Generative Proxemics: A Prior for 3D Social Interaction from Images. In *CVPR*, 2024.

Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. Multi-person Implicit Reconstruction from a Single Image. In *CVPR*, 2021.

Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. In *CVPR*, 2020.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *CVPR*, 2019.

Ulrich Pinkall and Konrad Polthier. Computing Discrete Minimal Surfaces and Their Conjugates. *Experimental Mathematics*, 1993.

Pablo Ruiz Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and Jose Garcia-Rodriguez. in2IN: Leveraging Individual Information to Generate Human Interactions. In *CVPRW*, 2024.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *ICCV*, 2019.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *CVPR*, 2020.

Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks. In *CVPR*, 2021.

Igor Santesteban, Miguel A. Otaduy, and Dan Casas. SNUG: Self-Supervised Neural Dynamic Garments. In *CVPR*, 2022.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human Motion Diffusion as a Generative Prior. In *ICLR*, 2024.

Zhengwentai Sun, Keru Zheng, Chenghong Li, Hongjie Liao, Xihe Yang, Heyuan Li, Yihao Zhi, Shuliang Ning, Shuguang Cui, and Xiaoguang Han. ReImagine: Rethinking Controllable High-Quality Human Video Generation via Image-First Synthesis. arXiv:2604.19720, 2026.

Pascal Volino and Nadia Magnenat-Thalmann. Resolving Surface Collisions through Intersection Contour Minimization. *ACM TOG*, 2006.

Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Cheng Lin, Rong Xie, Li Song, Xin Li, and Wenping Wang. Disentangled Clothed Avatar Generation from Text Descriptions. In *ECCV*, 2024a.

Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4D-DRESS: A 4D Dataset of Real-World Human Clothing With Semantic Annotations. In *CVPR*, 2024b.

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*, 2022.

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*, 2023.

Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. RigNet: Neural Rigging for Articulated Characters. *ACM TOG*, 2020.

Hang Ye, Xiaoxuan Ma, Hai Ci, Wentao Zhu, and Yizhou Wang. FreeCloth: Free-form Generation Enhances Challenging Clothed Human Modeling. In *CVPR*, 2025.

Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4D Instance Segmentation of Close Human Interaction. In *CVPR*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023a.

Weitian Zhang, Yichao Yan, Yunhui Liu, Xingdong Sheng, and Xiaokang Yang. E3Gen: Efficient, Expressive and Editable Avatars Generation. In *ACM MM*, 2024a.

Weitian Zhang, Sijing Wu, Manwen Liao, and Yichao Yan. Disentangled Clothed Avatar Generation with Layered Representation. In *ICCV*, 2025.

Xuanmeng Zhang, Jianfeng Zhang, Rohan Chacko, Hongyi Xu, Guoxian Song, Yi Yang, and Jiashi Feng. GETAvatar: Generative Textured Meshes for Animatable Human Avatars. In *ICCV*, 2023b.

Zechuan Zhang, Zongxin Yang, and Yi Yang. SIFU: Side-view Conditioned Implicit Function for Real-world Usable Clothed Human Reconstruction. In *CVPR*, 2024b.

Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. DeepMultiCap: Performance Capture of Multiple Characters Using Sparse Multiview Cameras. In *ICCV*, 2021a.

Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric Model-Conditioned Implicit Representation for Image-based Human Reconstruction. *IEEE TPAMI*, 2021b.

A Design Rationale: Applying the Pose in 2D versus 3D

Our pipeline makes two design choices: generating each person in 2D and lifting, rather than natively in 3D, and applying the target pose in 3D, rather than in the image plane. The first is a fidelity and text-control choice, settled by the generation FID (Sec. 4.2) and the qualitative comparison (Fig. 9). We treat the second here.

Once each person is generated as a 2D image, the target pose can be applied either in the image plane before lifting or in 3D after it. In the image plane there are two routes: generating directly in the target pose with a pose-conditioned image model (FLUX.1-dev with the ControlNet model `Shakker-Labs/FLUX.1-dev-ControlNet-Union-Pro-2.0` (Zhang et al., 2023a)), or reposing a canonical generation in 2D (ReImagine (Sun et al., 2026)). Both apply the pose to the image rather than to the 3D geometry, where depth and occlusion are unresolved, and neither yields a 3D asset that is both correctly posed and cleanly reconstructed.

Fig. 7 applies one target interaction pose to a single generated subject through the two image-plane routes and compares them with our 3D reposing. The pose-conditioned image model (FLUX with ControlNet) misreads the 3D pose. It renders a kneeling target as an upright torso, and even where the silhouette looks plausible it places the wrong foot forward, a depth-ordering error the 2D pose condition cannot encode. The 2D reposer (ReImagine) reaches the pose, but the posed image is self-occluding, so lifting it to 3D degrades the hidden regions. A frontal kneel occludes the rear foot and the lift returns distorted leg proportions with a mis-aligned kneeling angle, while a back occluded by hair reconstructs into a misshapen surface with a melted arm. Our route instead reposes a canonical, fully-visible asset, so neither the depth-ordering nor the occlusion failures arise.

One might supply the known target SMPL to a body-conditioned reconstructor to close the occlusion gap, but SMPL only fixes the pose, whereas the occluded garment appearance is unobserved and must be hallucinated, which current single-view reconstructors do not do reliably. A canonical, limb-separated pose avoids this. It has no self-occlusion, leaving only the usual unseen back. As a common pose, it is well-supported in training, so single-view reconstruction recovers it well. An arbitrary interaction pose has neither property. Recovering the hidden appearance directly, by generating extra views or requiring several reference images, is a heavier and different pipeline. We therefore reconstruct in the canonical pose and reach the target by reposing.

Even where a single posed subject is recovered well, composing two independently generated-and-reconstructed people into accurate contact is not free. Each arrives in its own camera and scale at a pose the generator only approximates, so the pair needs per-instance registration and a residual pose correction, effectively a second reposing. Reposing one canonical asset in 3D instead deforms a fixed surface directly to the target pose, so the two people share one frame at the correct relative placement without any separate registration, leaving only their interpenetration for the combination stage.

This is a trade-off. The 2D routes apply the pose in a single learned step and need no explicit geometry, but they hallucinate occluded structure and cannot be controlled to an exact contact. Our route keeps the operation geometric and controllable, at two costs. Our tightness-aware skinning (Sec. 3.3) resolves the first, the skinning artifacts of weight transfer. The second is inherent to kinematic reposing, which models no cloth physics, so wrinkles stay rigid and very loose cloth does not drape freely, as we discuss in App. E. Because our task is close interaction, where exact pose and contact are essential, we accept these costs and choose the controllable route.

B Tightness-Aware Skinning: Construction and Analysis

B.1 Background: smoothness energy and RSWT

The smoother Q of Eq. (2) is built from the cotangent Laplacian $L = D - \Omega$, with the standard cotangent edge weights $w_{ij} = \frac{1}{2}(\cot \alpha_{ij} + \cot \beta_{ij})$ that make it geometry-aware, assembled on the welded vertex set to avoid seam cracks. The Dirichlet term $W^\top L W$ interpolates but is only C^0 at the boundary. Adding the bi-Laplacian $\|LW\|^2_{M^{-1}}$ with the Voronoi mass matrix M gives the curvature-continuous fill $Q = L + LM^{-1}L$. Inpainting reliable boundary weights W_k over the rest



Figure 7: **Applying the pose in 2D versus 3D.** For two target interaction poses (top, front and side), the pose is applied by (ControlNet 2D) generating directly in the target pose with a pose-conditioned image model, (ReImagine 3D) reposing a canonical generation in 2D then lifting it, and (ours) reposing a single canonical 3D asset. The image-plane routes misread the 3D pose or degrade under self-occlusion after lifting, whereas our 3D reposing reaches the exact pose in a shared frame.

is the sparse solve $Q_{uu}\, W_u = -Q_{uk}\, W_k$. At $c \to 0$ our confidence solve reduces to exactly this inpaint.

RSWT (Abdrashitov et al., 2023) supplies the anchor weights our method starts from and defines the hard mask it generalizes. It accepts a cloth vertex as *matched* to its closest body point only within a distance and normal tolerance, copies weights there by barycentric interpolation, and inpaints the unmatched (loose) region with the same Q under a *hard* known/unknown mask, so the matched tight weights are frozen while Q shapes only the loose region. Our tightness-aware skinning (Sec. 3.3) replaces this hard freeze with a soft, per-vertex confidence.

B.2 Confidence-weighted weight solve

The per-vertex confidence $c(v)$ is a raised-cosine gate on the looseness,

$$c(v) = \tfrac{1}{2}\Big(1 + \cos\big(\pi\ \mathrm{clip}(\|\tau(v)\|\,/\tau_{\max},\, 0,\, 1)\big)\Big), \tag{8}$$

Table 4: **Reposing at full scale** (64 outfits $\times$ 32 target poses, 2,048 cases). The four distortion measures of Table 3 (strains in percent, E_{bend} in degrees), lower is better. Ours stays lowest on every measure, preserving the ordering and margins of Table 3 at scale.

Method	$\varepsilon_e \downarrow$	stretch$\downarrow$	$\varepsilon_a \downarrow$	$E_{\text{bend}} \downarrow$
naive	14.3	3.5	25.5	8.0
RSWT	11.5	2.8	19.1	5.5
GAPS	10.3	2.8	15.0	2.5
Ours	**8.8**	**2.1**	**12.7**	**1.44**

with cutoff $\tau_{\max} = 8$ cm, beyond which the tip anchor is treated as unreliable and the solve reduces to pure Q-smoothing. The gate only needs to fall off monotonically from tight to loose, so its exact form is not critical, and we use a raised cosine on $\|\tau\|$ for its smoothness.

Minimizing the confidence-weighted energy of Eq. (4) is a quadratic problem. With $C = \operatorname{diag}(c)$, setting the gradient to zero, $2\,C\,(W - W^0) + 2\lambda\,Q\,W = 0$, gives the normal equations $(C + \lambda Q)\,W = C\,W^0$ of Eq. (5). This system is sparse, symmetric, and positive-definite, and we solve it once by a direct sparse factorization. We use $\lambda = 10^{-3}$, and the raised-cosine confidence makes the reposing insensitive to its exact value. We then clamp negative entries to zero and renormalize each vertex's weights to sum to one, restoring a partition of unity.

B.3 Implicit multi-point averaging

Although each anchor $W^0(v)$ is borrowed from a single body point, the solve of Eq. (5) produces a multi-point blend. Writing the solution as $W = S\,W^0$ with $S = (C + \lambda Q)^{-1}C$, each final weight $W_i = \sum_k S_{ik}\,W^0_k$ is a weighted average of the anchors of nearby cloth vertices, with the weights summing to one since $Q\mathbf{1} = 0$ gives $(C + \lambda Q)\mathbf{1} = C\mathbf{1}$, hence $S\mathbf{1} = \mathbf{1}$. Because each of those anchors is itself a single body point, W_i becomes an implicit average over many body points even though each cloth vertex supplies just one. The confidence C controls how far this averaging reaches, with looser vertices diffusing more widely, while Q weights nearer anchors more. Crucially, the averaging happens *on the cloth surface*, blending the anchors of surface-adjacent vertices and thus respecting garment topology, whereas a body-space method such as GAPS (Chen et al., 2024) averages body points that are merely close in space and can blend weights across distinct body parts wherever those parts come close. Our solve thus attains this multi-point behavior while remaining garment-intrinsic and correspondence-aware. Empirically, adding GAPS's body-space averaging on top of our cloth-space solve leaves the reposing metrics essentially unchanged (ε_e $8.16 \to 8.08$), confirming that our solve already provides what it would add.

B.4 Reposing at scale

The main reposing benchmark (Table 3) uses 15 outfits, binned by looseness (five each), at 8 target poses each. To rule out sample-size effects, we repeat the evaluation on every clean-source outfit in 4D-DRESS, covering all 64 outfits (32 subjects, inner and outer layers) at 32 cross-sequence target poses each, a seventeen-fold larger set of 2,048 cases. The method ordering is unchanged and ours stays lowest on all four measures (Table 4), with the naive-to-ours bending ratio holding near $5.6\times$ from the 120-case main set up to the full 2,048-case one, so the result is not an artifact of the sample. Since 4D-DRESS has no perfectly limb-separated T-pose, each source here is its least-posed captured frame, leaving a small residual body pose that is identical across the four weight fields and does not affect their relative comparison.

C Penetration-Aware Combination: Construction and Analysis

C.1 Body Alignment

Before resolving the garments, we first separate the two SMPL bodies. Although the source interaction pose itself does not self-overlap, our reconstructed people have their own body shapes, so posing them into it can introduce new body penetration, which we clear so the garment solution starts from non-overlapping bodies. We seek the smallest per-person pose correction $\Delta = \{\Delta_1, \Delta_2\}$, one small

pose adjustment Δ_i per body (a change to body i's SMPL translation, orientation, and pose), that pushes each body out of the other:

$$\min_{\Delta} \sum_{v} \big[-s(v)\big]_+ \;+\; \gamma \sum_{i=1,2} \big\| V_i(\Delta_i) - V_i(\mathbf{0}) \big\|^2. \tag{9}$$

Here $s(v)$ is the signed distance of a vertex v to the other body (negative inside), so $[\,\cdot\,]_+ = \max(\cdot, 0)$ keeps only the penetrating vertices, those with $s(v) < 0$. $V_i(\Delta_i)$ is body i posed by the correction Δ_i, and $V_i(\mathbf{0})$ is the same body at its original interaction pose. The first, a penalty term, drives the bodies apart, while the second, a γ-weighted term, keeps the correction minimal so the interaction pose is preserved.

C.2 Cloth–Cloth Target-Surface Weighting

We ablate the confidence-weighted target surface X (Eq. (6)), $w = \|\tau_A\| / (\|\tau_A\| + \|\tau_B\|)$, against the plain equidistant midpoint ($w = \frac{1}{2}$). Both variants remove comparable overlap, with a residual volume that is marginally lower for the symmetric split (11.1 versus 13.1 cm^3 median). This aggregate is blind to *how* the overlap is split, which is what the weighting controls. Restricting to the displaced vertices, the confidence surface moves the looser half $1.23\times$ as far as the tighter half, against a near-even $1.01\times$ for the symmetric one, so the looser garment yields while the tighter keeps its fit. This is the physically correct direction and is consistent with the tightness signal used throughout, so we keep the confidence surface even though the aggregate volume does not reward it.

C.3 Displacement Solve

Negligible distortion. Our resolver barely distorts the cloth, a property that comes from the confidence-weighted solve and its boundary feathering analyzed below. Across the 900 stratified cases only a local region around each contact is displaced ($\sim 2.2\%$ of all garment vertices), and median edge-strain, area-strain, and bending are all ≈ 0, reaching only $0.72\%/1.78\%$ and $0.30°$ at the 95th percentile. The raw overlap also grows sharply with looseness (the loosest stratum enters with roughly $8\times$ that of the tightest, $\sim 4{,}200$ vs. $\sim 500\,\mathrm{cm}^3$), yet the 92–96% reduction of Table 5 holds in every stratum, so it is not carried by the easy, tight cases.

Why the confidence-weighted solve. A simpler resolver would move each penetrating vertex straight onto the interface by its target d^0, with no solve. This displaces only the penetrating vertices while their neighbours stay fixed, so the surface breaks into a crease where the moved patch meets the untouched cloth. Our confidence-weighted solve instead assigns low confidence to the loose cloth around the patch, where nothing penetrates ($d^0 = 0$). Its smoothness term then diffuses the correction outward from the penetrating patch across these low-confidence vertices, so the displacement fades to zero over the surrounding cloth rather than stopping abruptly at the patch edge and tearing it. Across a stratified set of 30 interactions the two remove comparable overlap, but the raw move stretches the cloth far more, its worst-case edge stretch reaching $1.6\times$ against $1.1\times$ for our solve, and $2.1\times$ on the highest-overlap interactions (Table 6).

Why the correction is feathered. Within the solve, the penetration target d^0 is nonzero only on the penetrating patch, unlike the skinning anchors that are defined at every vertex and vary smoothly with tightness, so its boundary is a sharp discontinuity between the pushed and the untouched cloth. Feathering the confidence to zero over a few rings there keeps that boundary crease-free, and it is the one ingredient the skinning solve does not need. Without it the sharp boundary leaves a crease that the iterated solution compounds into a growing flap, raising the worst-case edge stretch to $1.6\times$ against $1.1\times$ for ours (Table 6). This aggregate matches the no-solve case, but the failure differs in kind, a localized flap rather than the broad shred left without the solve (Fig. 8).

Self-penetration and iteration. By construction, the same interface and solve apply to *self-penetration* as well, where one person's garment crosses itself or its own body, a case reposing can leave behind. Only the interface orientation differs, so the single formulation of Eq. (7) covers penetration both within and between the two people. The pipeline resolves both, though we evaluate the interaction-specific inter-penetration quantitatively. Separately, because each pass under-shoots slightly and can open a fresh overlap elsewhere, we iterate the solve, rebuilding the interface X each time and keeping the pass with the least residual overlap.

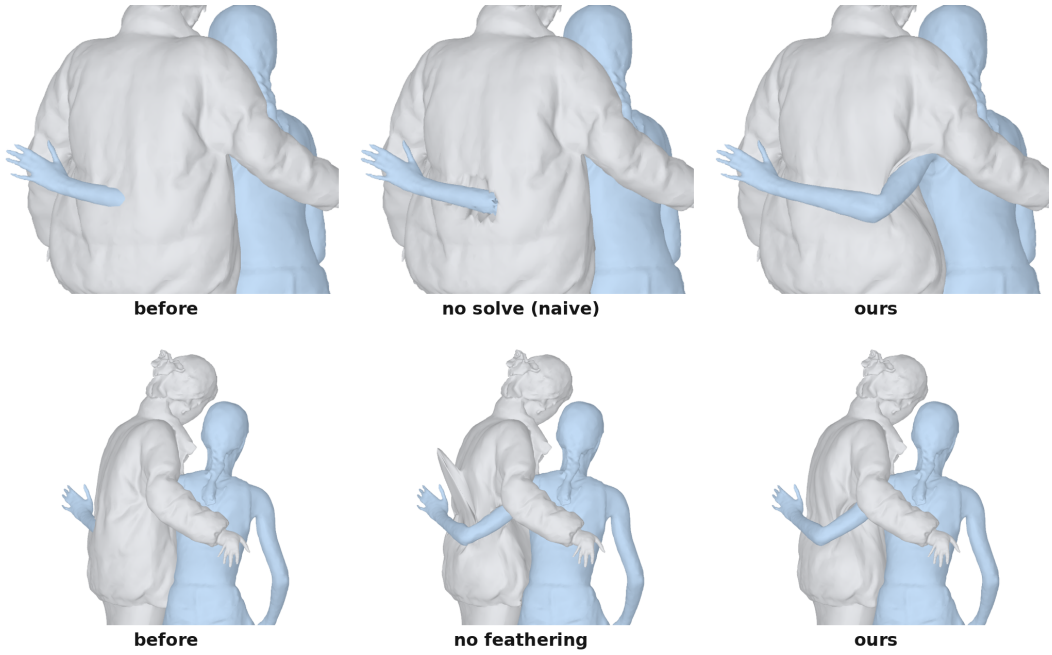


Figure 8: **Ablating the penetration solution.** An interacting pair whose garments interpenetrate where one arm wraps over the other's back. *Top row* (contact region): moving each penetrating vertex straight onto the interface with no solve shreds the cloth at the contact, while our confidence-weighted solve separates the garments cleanly. *Bottom row* (a different view): dropping the boundary feathering lets the sharp patch boundary compound into a growing flap over the iterated passes, which our feathering removes. Each row shows the raw *before* state at left. Both ablations leave cloth-level artifacts that ours avoids.

Table 5: **Cloth-cloth interpenetration**, multi-outfit (900 cases, six looseness strata from LL to TT pairing loose (L), medium (M), and tight (T) garments, random outfit per case). Percentage of the raw overlap volume removed by the take-best resolver, robust across looseness.

	LL	LM	LT	MM	MT	TT	All
interpen. reduction ↑	96%	94%	93%	93%	92%	93%	**94%**

D Additional Generation Evaluation

D.1 FID protocol and reference selection

Evaluating generation quality is less standard here than reposing or interpenetration, since a freely generated clothed human has no paired ground truth to compare against. The usual practice of scoring a generator against its own training distribution does not transfer to our setting. Our pipeline is training-free, the baselines are each trained on different data, and the task has no shared target distribution, so scoring each method against its own training distribution would reward whichever model best reproduces its own data rather than whichever produces the better clothed human. We instead evaluate every method against one common held-out real reference, as 3D human generators are scored (Dong et al., 2023), applied identically to all, so the comparison is not tailored to any method's own distribution. Each method is shown through its own native renderer at a fixed set of views, following standard practice for 3D generation evaluation, which avoids the conversion artifacts that forcing every representation through a single renderer would introduce and that would unfairly penalize the Gaussian baselines. No single real capture carries both photoreal appearance and true 3D geometry, so we use two reference modalities and pair each metric with the ground truth it can actually measure, using photographs (SHHQ, FlickrCI3D) for appearance and 3D scans (4D-DRESS, Hi4D) for geometry. This is why appearance and geometry are reliably measured against different references throughout Sec. 4.2. To keep these distances a measure of generation quality

Table 6: **Penetration-solution ablation** on a stratified set of 30 interactions. All three variants remove comparable interpenetration, but dropping the confidence-weighted solve (*no solve*, a direct push) or its boundary feathering (*no feathering*) distorts the cloth far more. We report the worst-case edge stretch (the 99.5th-percentile edge elongation of Sec. 4.1), overall and split by overlap severity. Lower is better.

	stretch↓	stretch by overlap		
		low	med	high
no solve	1.63	1.03	1.76	2.10
no feathering	1.61	1.15	1.67	2.00
ours	**1.09**	**1.03**	**1.07**	**1.15**

Table 7: **Front-facing pseudo-normal geometry on photo datasets** (pseudo-normal from Sapiens). SHHQ (single-human, $N = 5{,}000$) reports FID and KID, and FlickrCI3D (two-person, $N = 1{,}281$) reports KID only. KID $\times 10^{-3}$, lower is better, **best** and second.

	SHHQ (single)		FlickrCI3D (2p)
Method	FID↓	KID↓	KID↓
GETAvatar	47.7	41.5	36.9
GSM	39.2	39.0	31.5
E3Gen	35.5	29.2	**23.2**
LayerAvatar	44.6	42.7	49.7
Ours	**23.9**	**18.1**	23.9

rather than of pose, every method is evaluated at the same pose distribution as the reference, with the poses fit from the reference images, on the resulting posed output rather than a canonical one. Because FID is biased at finite sample sizes, we recompute all methods under this single protocol, report KID alongside FID, and read only the relative ordering across methods rather than absolute scores, which is also why we do not cite the numbers reported in the baselines' own papers.

D.2 Front-facing pseudo-normal geometry

Geometry FID is computed on the surface normals of both the generated result and the reference. On the scan datasets (4D-DRESS, Hi4D) the reference has true normals, and the main paper compares each method's own native normals against them, except GSM, whose Gaussian representation yields no native surface normal (shown as "–"). The photo datasets, SHHQ (single) and FlickrCI3D (two-person), have no ground-truth normal, so their reference is estimated with an off-the-shelf predictor (Sapiens (Khirodkar et al., 2024)) in either case. The main paper uses each method's native normal for the generated result. Here we instead estimate that normal the same way, from a front-view render, so the generated result and the reference pass through one predictor. This covers every method, GSM included, and removes the native-versus-estimated asymmetry. Because Sapiens is trained predominantly on frontal imagery, this pseudo-normal is reliable only for the front-facing single-view crops these photo datasets provide, and on the multi-view scan datasets, where side and back views dominate, it degrades and we do not use it.

On the single-human photographs (SHHQ) our method attains the lowest FID and KID by a wide margin (Table 7), consistent with the scan-referenced geometry result in the main paper. On the two-person photographs (FlickrCI3D) our pairs are a close second in geometry, mirroring the near-tie on Hi4D scans reported in the main paper. For FlickrCI3D, whose sample is small, we report KID only (unbiased at small sample sizes), as in the main paper.

D.3 Full single-human qualitative comparison

Fig. 9 shows the qualitative single-human generation across the full baseline set of Table 1, together with the text-to-3D generator HumanGaussian (Liu et al., 2024). The unconditional generators cannot follow a prompt and return an arbitrary outfit per sample. GSM renders sharp texture, having been trained on 2D human photographs, but carries no surface normal. HumanGaussian does follow the prompt, yet its results are cartoonish and low-fidelity, and it too has no surface normal. Only our



Figure 9: **Single-human generation: appearance and geometry.** Each method is shown as texture and native surface normal (the two axes of Table 1) for three target poses. The rows are the four unconditional generators, the text-to-3D HumanGaussian, and our text-conditioned lifting. GSM and HumanGaussian provide no surface normal, and only our result stays sharp in both texture and normal.

text-conditioned lifting both follows the prompt and stays sharp in appearance and geometry across all poses.

D.4 Additional two-person results

Fig. 10 extends the two-person gallery of Fig. 4. Because the two people are generated independently and only assembled at the end, any pair of single-person prompts can be combined, so the reachable interactions grow combinatorially with the per-person variety. As in the main figure, no baseline is shown, since no prior method generates interacting clothed pairs.

E Limitations

Our pipeline is training-free and assembles off-the-shelf models through geometric steps. This is what keeps it controllable, but it bounds the results in two ways. First, it models no cloth physics, so it cannot reproduce how cloth deforms under gravity and contact. Second, its quality inherits that of the off-the-shelf models it builds on, whose errors compound through the stages.

In *reposing*, because a garment is generated once and carried by skinning, no pose-dependent wrinkles form, and existing wrinkles move rigidly with the limbs. Very loose cloth is the hardest case,



Figure 10: **Two-person generation gallery (extended).** Additional interacting clothed pairs generated by our pipeline, extending Fig. 4. Front view.

as it lies far from the body where the tightness correspondence is weak, so its skinning weights are smoothed from the tighter regions closer in and it is carried with the limbs. Our tightness-aware skinning keeps such loose regions from collapsing onto the body, but it does not restore free drape. In *combination*, the resolver removes the interpenetration between the two garments, but two cases fall outside it. A hand can still pass through the other person's garment, which the resolver does not target, since it separates overlapping garments rather than a body passing through cloth. And because it optimizes only to remove overlap, with no term that preserves contact, separating one can open a small gap where the garments should stay touching. Fig. 11 shows representative cases.

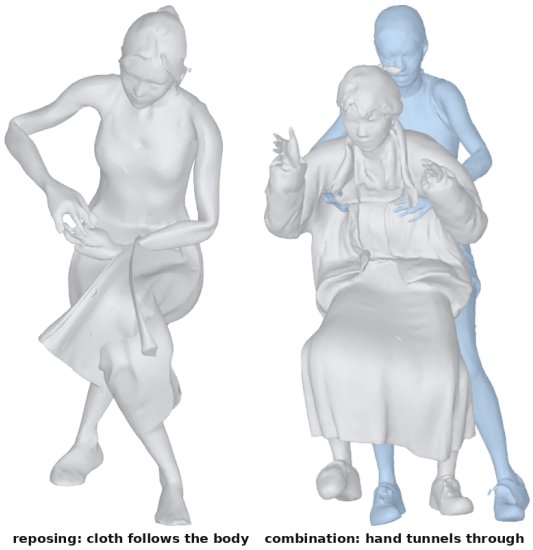


Figure 11: **Representative failure cases.** (left) In reposing, loose cloth is carried rigidly by the skinning and follows the body instead of draping freely and re-forming under the pose. (right) In combination, a hand can tunnel through the other person's garment, a cloth-body penetration our cloth-cloth resolver does not target.

None of these break the interaction structure. The target pose and the contact placement are still met, and the failures are confined to cloth-level detail. The natural remedy is to layer a cloth simulator on our geometric assembly, which would introduce pose-dependent wrinkling and free drape and, with a contact term, both stop the hand tunneling and hold intended contact. We leave this integration to future work.