NATIVE 3D HUMAN GENERATION WITH GEOMETRY-ALIGNED SOFT RIGGING FROM A SINGLE IMAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating high-fidelity, riggable 3D human avatars from a single image remains a fundamental challenge. Native 3D models suffer from limited data and poor generalization. Meanwhile, 2D-inspired methods rely on multi-view or video supervision, which introduces cross-view inconsistency and unstable convergence. To address these limitations, we present NVWA, a native 3D generation framework that reconstructs geometry-aligned, riggable avatars from one image. Our method makes two key innovations. First, rather than supervising with multi-view renderings or Internet videos, we synthesize large-scale explicit 3D human assets for direct native 3D supervision. This eliminates the multi-view inconsistency inherent in distillation-based methods and enables substantial gains in geometric fidelity. Second, we depart from predefined parametric topologies (e.g., SMPL). Such templates constrain outputs to naked-body shapes and cannot model loose clothing or accessories. Instead, we propose soft rigging. Our model jointly generates free-form volumetric geometry and a spatially aligned semantic correspondence field. This field establishes per-vertex correspondence to a standard rigging template without restricting the generated shape. Shape generation is thus decoupled from template constraints while full pose controllability is preserved. To support training, we introduce a scalable pipeline for synthesizing photorealistic explicit 3D humans. We also release NVWA-100K, a dataset of 100,000 high-fidelity textured mesh models. Experiments demonstrate that NVWA achieves state-of-the-art fidelity and rigging quality, significantly advancing single-image 3D human generation. Code and additional qualitative results are available in an anonymous repository.

1 INTRODUCTION

3D human models play an increasingly central role in fields ranging from human–computer interaction and gaming to digital content creation. In recent years, large-scale 3D generation models have substantially improved 3D modeling efficiency, introducing transformative technologies and gaining widespread attention. Compared with generic 3D object generation, human modeling imposes stricter requirements on geometric fidelity and riggability. These remain unresolved challenges for existing 3D human modeling approaches.

Prior 3D human generation models fall into two broad paradigms. The first category is native 3D generative models, further divided into general 3D generative models and human-specific ones. General 3D generative models (Xiang et al., 2026; Wu et al., 2025) struggle to produce photorealistic 3D humans and lack riggability. This is primarily because the high acquisition costs of realistic 3D human data lead to biased training sources – few photorealistic human models but more cartoonish models. Although human-specific 3D generative models (Zhuang et al., 2025; Qiu et al., 2025) avoid data bias, they are constrained by limited 3D data, leading to suboptimal generalization and refinement. The second category is 2D-inspired methods, which leverage powerful 2D generative models to train high-quality 3D human generative models. Score Distillation Sampling (SDS) (Poole et al., 2022) serves as a representative instance of this methodology. Nevertheless, these approaches have significant limitations in maintaining multi-view consistency, leading to unstable convergence and inefficient 3D model reconstruction.

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.

In contrast, we propose a native 3D generation framework tailored for digital avatars, centered on two core ideas: 1) generating large-scale 3D model data, and 2) creating an explicit 3D representation to support precise human pose rigging and appearance rendering. IDOL (Zhuang et al., 2025) and LHM (Qiu et al., 2025) pioneered this direction and remain strong state-of-the-art models. Our method advances this line of work with two key innovations:

First, rather than relying on multi-view data or Internet videos for training, we construct explicit, photorealistic 3D human assets to directly supervise native 3D generation. Specifically, IDOL synthesized large-scale multi-view data for training, but multi-view generation inherently produces inconsistent content, severely impacting quality. LHM utilized internet-sourced data, but its human parametric model fitting process introduced inherent semantic misalignments between video frames. These limitations cap the performance of 3D avatar generation models trained on multi-view or video data. Instead, we develop a method to synthesize large-scale explicit 3D human models, constructing a massive explicit dataset of 3D mesh models. This eliminates the multi-view inconsistency intrinsic to 2D-inspired methods, enabling breakthroughs in modeling refinement.

Second, we discard the strategy of generating geometric appearances on pre-defined UV topologies (e.g., SMPL-based parametric models)—adopted by E3GEN (Zhang et al., 2024), IDOL, and LHM, in favor of a "free-form geometry generation together with soft rigging" approach. While predefined UV topologies simplify generation via semantic alignment, they constrain results to naked bodies and fail to accurately model loose clothing, backpacks, hats, or accessories. Thus, we decouple parametric model constraints during generation, enabling free-form volumetric modeling, and reconstruct rigging in a subsequent stage. We achieve this through per-vertex semantic correspondence generation, which produces semantic labels alongside geometry and appearance, establishing a "soft" correspondence between the generated model and a standard rigging template. Experiments show that this approach improves generation precision and significantly enhances pose-rigging quality.

Our contributions can be summarized as:

• **Native 3D human generation.** We adapt a native 3D generative prior to clothed humans using large-scale explicit 3D assets, avoiding the cross-view inconsistencies associated with independently synthesized multi-view supervision and improving fidelity on real-world human images.

• **Soft rigging.** We propose a soft rigging strategy that simultaneously generates human geometry and a spatially aligned, template-decoupled semantic correspondence field, rather than generating on a predefined topology. This decouples geometric generation from template constraints, enabling expressive modeling of loose garments and accessories while supporting riggability.

• **Scalable explicit human data.** We present a scalable data acquisition pipeline for generating photorealistic 3D human assets, accompanied by the release of NVWA-100K – a large-scale dataset comprising 100K high-fidelity human models, achieving one of the largest collections of explicitly represented textured 3D humans.

2 Related Work

2.1 3D Human Generation

Early human reconstruction methods recover geometry from a single image (Zheng et al., 2019), monocular video (Peng et al., 2021), or multi-view observations (Işık et al., 2023). They commonly employ implicit or explicit surface representations (Saito et al., 2019; 2020; Xiu et al., 2022; 2023), together with parametric-body (Loper et al., 2015; Pavlakos et al., 2019) or pixel-aligned priors. Although effective for human-specific reconstruction, these approaches do not directly provide a scalable generative prior for diverse appearances.

With the rapid progress in generative models (Karras et al., 2019) and neural rendering (Mildenhall et al., 2021), a line of work (Abdal et al., 2024; Chen et al., 2023; Hong et al., 2022; Xiong et al., 2023; Zhang et al., 2023; Dong et al., 2023) explored synthesizing human bodies from unstructured 2D datasets (Fu et al., 2022) via 3D-aware GANs (Chan et al., 2021; 2022). However, due to the limited scale and diversity of such datasets suitable for adversarial training, these methods often struggle to produce realistic human geometry and appearance. More recently, diffusion-based

approaches with dataset scaling (Schuhmann et al., 2022) have enabled more diverse human generation (Kolotouros et al., 2023; Liu et al., 2024; Tang et al., 2024b; Zhou et al., 2024b) by leveraging score distillation losses (Poole et al., 2022) from pretrained diffusion models (Rombach et al., 2022), though they may still lack photorealism and consistency. In parallel, multi-view diffusion models (Chen et al., 2025) have been introduced to generate multi-view images of a given subject, which are then converted into 3D meshes via post-processing (Li et al., 2025). While these methods benefit from strong 2D appearance priors, inconsistencies among generated views can affect their geometry, especially in occluded regions and articulated body parts. In contrast, native 3D generators such as TRELLIS.2 (Xiang et al., 2026) and Pixal3D (Li et al., 2026a) directly model 3D assets but do not provide human-specific rigging semantics.

Animation-ready methods introduce parametric-body structure during generation. IDOL (Zhuang et al., 2025) predicts Gaussian attributes in a predefined SMPL-X (Pavlakos et al., 2019) UV space, while LHM (Qiu et al., 2025) conditions reconstruction on SMPL-X geometry. Such representations facilitate animation but couple the generated asset to a body template, making loose garments and accessories more difficult to represent. NVWA instead generates free-topology geometry together with a geometry-aligned semantic UV field, combining free-form surface modeling with the correspondence required for subsequent SMPL-X fitting.

2.2 3D Human Datasets

High-quality 3D human data constitute a fundamental resource for learning realistic human geometry and appearance. Existing datasets fall into three broad categories. First, scanner-based datasets, including 2K2K (Han et al., 2023), THuman (Su et al., 2022; Yu et al., 2021; Zheng et al., 2022), CustomHumans (Ho et al., 2023), and commercial scan collections such as Twindom, provide explicit meshes with high-quality local geometry and textures. However, acquiring and manually processing real scans is expensive, which limits the number of identities, clothing styles, and poses that can be captured.

Second, depth-sensor (Liu et al., 2019; Shahroudy et al., 2016) and multi-view capture datasets (Fang et al., 2021; Teufel et al., 2025; Zheng et al., 2021) collect human motion using RGB-D cameras or synchronized camera systems. Datasets such as ZJU-MoCap (Peng et al., 2021), HuMMan (Cai et al., 2022), DNA-Rendering (Cheng et al., 2023), and MVHumanNet (Xiong et al., 2024) contain large numbers of frames and support dynamic reconstruction and novel-view synthesis. Nevertheless, their apparent scale is often measured in frames rather than distinct identities or reusable 3D assets. Moreover, many of these datasets provide calibrated multi-view images, videos, or fitted parametric bodies rather than diverse, independently textured explicit meshes. Consequently, their data distribution is dominated by a comparatively limited set of subjects, garments, and capture environments.

Third, recent synthetic datasets improve scalability by generating human images across diverse identities and viewpoints. HuGe100K, for example, provides large-scale multi-view renderings of generated humans and substantially increases identity diversity. However, multi-view image collections are not equivalent to explicit 3D asset datasets: independently generated views may contain geometric or appearance inconsistencies. To bridge the gap between the fidelity of real scans and the scalability of synthetic image collections, we construct NVWA-100K, a large-scale dataset containing 100K textured explicit human meshes.

3 Method

In this section, we introduce NVWA, a framework that generates a textured, free-topology human mesh and a geometry-aligned soft rigging representation from a single image. As illustrated in Fig. 1, Sec. 3.1 describes the construction of NVWA-100K. Sec. 3.2 introduces the soft rigging representation and the subsequent SMPL-X fitting and skinning-weight transfer procedure. Sec. 3.3 presents the native 3D generation framework, and Sec. 3.4 describes image-space A-pose editing for animation.

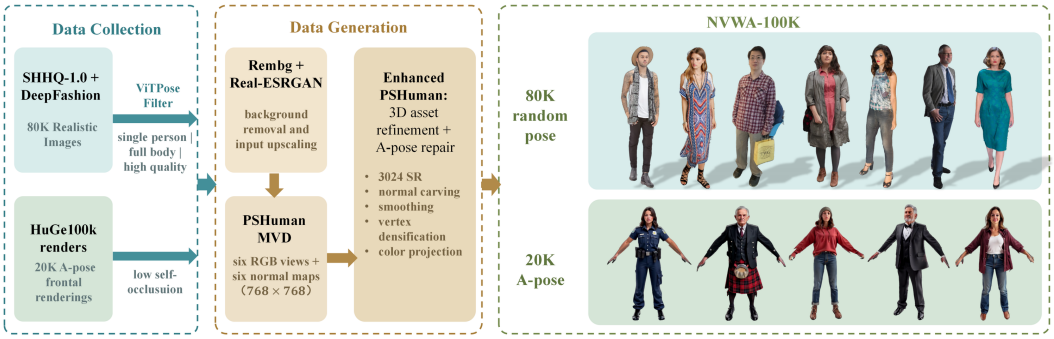

Figure 1: **Scalable construction of NVWA-100K.** We present a scalable data acquisition pipeline for generating photorealistic 3D human assets.

3.1 Scalable Explicit 3D Human Data

Inspired by generic 3D generation models, we observed that a large-scale, high-quality explicit 3D human dataset is essential for building 3D human generation models. In this section, we introduce an efficient pipeline for generating 100K photorealistic 3D humans.

Data Collection. For the 80K random pose human models, we curate high-quality realistic human images from the SHHQ-1.0 (Fu et al., 2022) and DeepFashion (Liu et al., 2016) datasets. We use ViTPose (Xu et al., 2022) to assess person count and body visibility, retaining high-quality, full-body images containing a single person. For the 20K A-pose 3D humans, we selected frontal renderings of 20K distinct identities from the HuGe100K dataset provided by IDOL (Zhuang et al., 2025). Because self-occlusion is least pronounced in A-pose human models, this additional 20K subset adds notable value for improving 3D human modeling.

Data Generation. As illustrated in Fig. 1, we first used Rembg (Qin et al., 2020) to remove the backgrounds, followed by upscaling the background-removed images using Real-ESRGAN (Wang et al., 2021). These enhanced images were fed into the pre-trained MVD of PSHuman (Li et al., 2025), generating six RGB images and normal maps from predefined viewpoints for each identity, all at a resolution of 768. To further improve quality, we upscaled both the images and normal maps to 3024 using Real-ESRGAN. We applied the explicit human carving strategy from PSHuman to generate initial 3D human models based on the upscaled normal maps. To refine the human models, we applied Laplacian smoothing to reduce surface artifacts. In addition, we incorporated the appearance projection strategy from PSHuman, mapping the colors from six predefined viewpoints' RGB images onto corresponding human model vertices by a vertex-weighted coloring method. Before applying appearance projection, we significantly increased the number of human model vertices, a modification that substantially improved the quality of the human model obtained through the vertex coloring strategy.

3.2 Geometry-Aligned Soft Rigging

Soft rigging representation. We generate a free-topology human mesh together with a per-vertex semantic correspondence field referenced to the canonical SMPL-X 2D UV atlas (Pavlakos et al., 2019). The generated vertices and faces are not constrained to follow the SMPL-X topology; only the correspondence field uses SMPL-X as its semantic reference. We call this representation "soft rigging" because it provides template-referenced correspondence on a freely generated surface, rather than imposing the template topology on the generated geometry:

$$M = (X, F, U) \tag{1}$$

where X and F define the mesh geometry, $U = \{\tilde{\mathbf{u}}_i\}_{i=1}^{N}$ is the three-channel soft rigging attribute replacing the vertex color of conventional texture meshes, encoding correspondence to the template. For each clothed human asset $M^h = (X, F)$, with $X = \{\mathbf{x}_i\}_{i=1}^{N}$, we use EVA-Tracker (Zhang et al., 2026) to obtain a fitted SMPL-X mesh $M^s = (Y, F_s)$, $Y = \{\mathbf{y}_j\}_{j=1}^{10475}$. We assign each SMPL-X vertex a fixed semantic UV coordinate $\mathbf{u}_j^s = (u_j^s, v_j^s) \in [0, 1]^2$ from the predefined 2D UV space

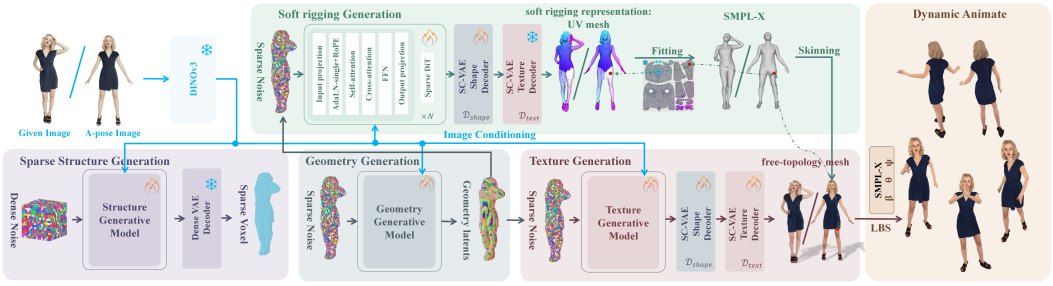

Figure 2: **Overview of NVWA.** Given a single image, NVWA simultaneously generates a free-topology human with texture and a soft rigging representation.

of the SMPL-X models, which is shared across identities and poses. For each clothed human asset vertex, we transfer the coordinate of its nearest fitted SMPL-X vertex:

$$k_i = \arg\min_j \|\mathbf{x}_i - \mathbf{y}_j\|_2^2, \qquad \mathbf{u}_i^h = \mathbf{u}_{k_i}^s. \tag{2}$$

Writing $\mathbf{u}_i^h = (u_i^h, v_i^h)$, we encode the field as a three-channel pseudo-color attribute, $\tilde{\mathbf{u}}_i^h = (2u_i^h - 1,\ 2v_i^h - 1,\ 1)$. This encoding enables semantic UV coordinates to be processed directly by the existing texture-generation pipeline. For storage and rendering, the attributes are linearly mapped from $[-1, 1]$ to $[0, 255]$.

Soft rigging-guided fitting and skinning. At inference time, the predicted rigging attributes provide correspondence for fitting a parametric body to the generated mesh. For each generated vertex $\hat{\mathbf{x}}_i$, we retrieve the nearest SMPL-X vertex in semantic UV space:

$$a_i = \arg\min_j \|\hat{\mathbf{u}}_i^h - \mathbf{u}_j^s\|_2^2. \tag{3}$$

The resulting correspondence set $C = \{(i, a_i)\}_{i=1}^N$ remains fixed during fitting. Let $\mathbf{v}_j(\Theta)$ denote SMPL-X vertex j under parameters Θ, including body shape, body and hand poses, global orientation, translation, and scale. We optimize a weighted geometric fitting term:

$$\mathcal{L}_{\text{data}}(\Theta) = \frac{1}{Z}\sum_{i=1}^{N} m_i w_i \left\|\hat{\mathbf{x}}_i - \mathbf{v}_{a_i}(\Theta)\right\|_2^2, \qquad Z = \sum_{i=1}^{N} m_i w_i, \tag{4}$$

where m_i is an inlier indicator and w_i is a body-region weight. We supplement this term with anatomical priors and refine the fit using silhouette-containment constraints in three orthographic views. In our implementation, we optimize the resulting fitting objective with L-BFGS and a strong-Wolfe line search; the full fitting procedure takes approximately 25 seconds per avatar on an NVIDIA RTX 3090 GPU. Finally, we transfer skinning weights from the fitted SMPL-X body to the generated mesh using the resulting correspondence set and animate it through linear blend skinning, achieving semantically robust weight transfer on free-form clothed human topology. More details can be found in Appendix A.1.

3.3 Native 3D Human Generation

Generation backbone. NVWA builds on the O-Voxel representation and structured latent generation pipeline of TRELLIS.2 (Xiang et al., 2026). Starting from image-conditioned sparse structure and geometry latents, two geometry-conditioned branches predict appearance and soft rigging attributes, respectively. The rigging branch follows the same generation paradigm as the texture branch, but predicts a soft rigging attribute instead of appearance features. NVWA therefore generates geometry, along with geometry-conditioned appearance and rigging attributes, within a unified pipeline, rather than generating an avatar on a predefined SMPL-X topology.

Human-domain adaptation. To transfer the generic 3D prior to the distribution of clothed humans without the prohibitive cost and potential prior forgetting, we fine-tune from TRELLIS.2's checkpoint rather than training from scratch. Specifically, we fully fine-tune the sparse-convolutional SC-VAE and apply LoRA (Hu et al., 2021) to the DiT models. For the soft rigging generation

Table 1: Quantitative comparison on X-Humans and in-the-wild images. **Bold** and underlined values indicate the best and second-best results in each column. "–" denotes an unavailable result.

	X-Humans					In-the-wild			
Method	**CD↓**	**ULIP-2↑**	**Uni3D↑**	**CLIP↑**	**LPIPS↓**	**ULIP-2↑**	**Uni3D↑**	**CLIP↑**	**LPIPS↓**
DreamGaussian	–	–	–	0.7739	0.1464	–	–	0.7641	0.3244
LGM	–	–	–	0.7770	0.1362	–	–	0.7819	0.3257
TRELLIS.2	2.03	0.3610	0.3042	0.8481	0.0827	0.3693	0.3298	0.8137	0.3522
Pixal3D	2.12	0.3360	0.2822	0.8529	0.0856	0.3780	0.3331	0.8365	0.3646
Human3Diffusion	–	–	–	0.8398	0.1218	–	–	0.7887	0.3307
PSHuman	3.25	**0.3976**	0.3182	0.8415	**0.0762**	0.3695	0.3305	0.8153	0.3174
IDOL	–	–	–	0.8311	0.1385	–	–	0.7698	0.3140
LHM	–	–	–	0.8341	0.1233	–	–	0.7789	0.3382
Ours	**1.74**	0.3638	**0.3190**	**0.8749**	0.0813	**0.3785**	**0.3335**	**0.8687**	**0.3108**

branch, we inherit the same DiT architecture as TRELLIS.2's texture generation branch. We first train on the large-scale synthetic NVWA-100K dataset to learn a comprehensive human prior.

Real-Scan Post-Training. The scale and diversity of NVWA-100K support broad human-domain coverage, but its generated assets can contain over-smoothed surfaces and residual reconstruction artifacts. We therefore introduce post-training using real-world scans, which provide finer local geometry and appearance details. More details are provided in Appendix A.2.

3.4 Image-Space A-Pose Editing for Animation

High-quality static geometry does not necessarily yield reliable animation. In-the-wild images often contain arbitrary poses and self-occlusion, which may introduce spurious connections in the mesh. Such connections can be stretched, causing adhesion, severe distortion, and surface tearing under large articulations. To mitigate this topology ambiguity before 3D generation, we use GPT-Image2 as an off-the-shelf image editor to obtain an identity- and clothing-preserving A-pose image with spatially separated limbs. Importantly, image-space A-pose editing does not establish a canonical 3D pose or replace parametric fitting: the generated mesh still undergoes SMPL-X fitting and skinning-weight transfer as described in Sec. 3.2. Additional details are provided in Appendix A.3.

4 Experiments

4.1 Implementation details

Our NVWA model is trained on a cleaned subset of NVWA-100K with an aesthetic score over 5.0. We render 16 views per human mesh at 1024×1024 resolution as conditional images. We train the model on 8 NVIDIA A6000 GPUs using AdamW with a learning rate of 5×10^{-4}. The LoRA configuration contains 0.089B trainable parameters, corresponding to 6.9% of full fine-tuning. Appendix A.2 provides the complete training configuration.

We use 50 in-the-wild images in visual results, qualitative and quantitative comparisons, and user studies, which are not included in our training data. For another quantitative analysis, we utilize 100 scans from 20 subjects, with 5 frames randomly selected per subject from X-Humans (Shen et al., 2023), which is not in our training data.

4.2 Comparison with Existing Methods

Comparison on Single-image Reconstruction. We compare NVWA with both general 3D generative models and human-specific models. Fig. 3 and Fig. 4 show comparisons with general and human-specific 3D generative models, respectively. We conduct the comparisons on challenging in-the-wild images featuring diverse poses, clothing styles, and body shapes. Tab. 1 reports quantitative results on the X-Humans dataset and in-the-wild images. For X-Humans, we evaluate 100 scans from 20 subjects, with five frames randomly sampled per subject. Following prior work, Chamfer Distance (CD) values are scaled by 10^4 for readability. The CLIP score (Radford et al., 2021) measures image-conditioned semantic consistency, while ULIP-2 (Xue et al., 2024a) and Uni3D (Zhou et al., 2024a) assess geometric similarity. For X-humans with gt geometry, we additionally evaluate



Figure 3: Qualitative comparison with general 3D generative models on in-the-wild images.

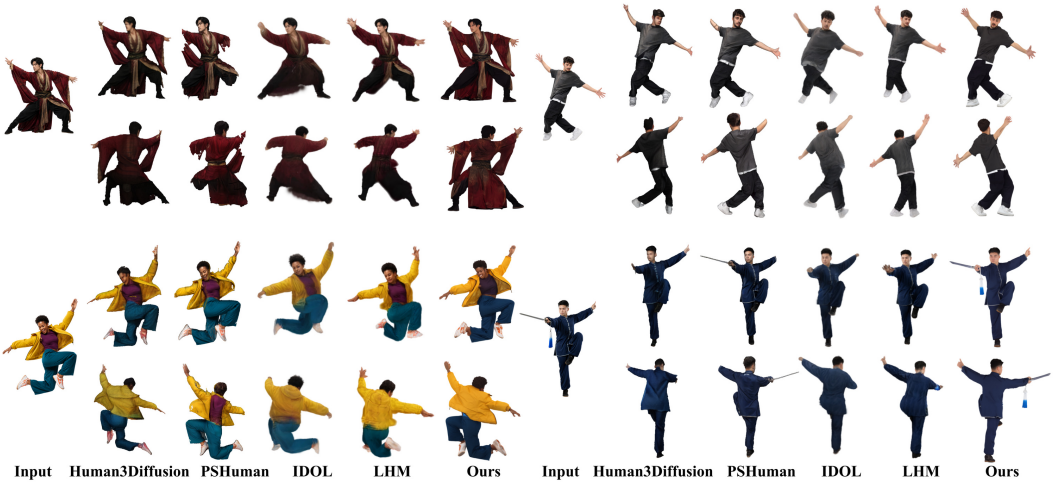

Figure 4: Qualitative comparison with human-specific models on in-the-wild images.

Chamfer Distance (CD) on point clouds sampled from visible surfaces and assess surface quality metrics using LPIPS (Zhang et al., 2018) of rendered RGB maps. Additional qualitative results are provided in Appendix A.5, and evaluation protocols are detailed in Appendix A.4.

Among the compared methods, DreamGaussian (Tang et al., 2024b) is an SDS-based model, LGM (Tang et al., 2024a) is a feed-forward model, while TRELLIS.2 (Xiang et al., 2026) and Pixal3D (Li et al., 2026a) are 3D diffusion models. These general 3D models, while producing plausible overall shapes, often fail to capture fine-grained details, generating blurred or twisted faces, and their outputs are not readily animatable.

Among human-specific methods, Human3Diffusion (Xue et al., 2024b) and PSHuman (Li et al., 2025) adopt MVD and PSHuman achieves high quality by scaling up MVD resolution. However, their outputs are still unrigged. In contrast, compared with the two pioneers in rigged 3D human generation, IDOL (Zhuang et al., 2025) and LHM (Qiu et al., 2025), our approach produces more accurate geometry, particularly in loose clothing. Unlike the SMPL-bounded methods, the design of NVWA allows for a flexible surface. Further, while it does not enhance facial features with face crops like PSHuman and LHM, NVWA produces clearer faces. Clothing texture is also more detailed, contributing to the overall higher visual quality of our method.

Comparison on Human Animation. The soft rigging generation method of NVWA is assessed by animating the human avatars to given poses. We use 20 SMPL-X motion sequences from AMASS (Mahmood et al., 2019) covering various activities including walking, dancing, and sports

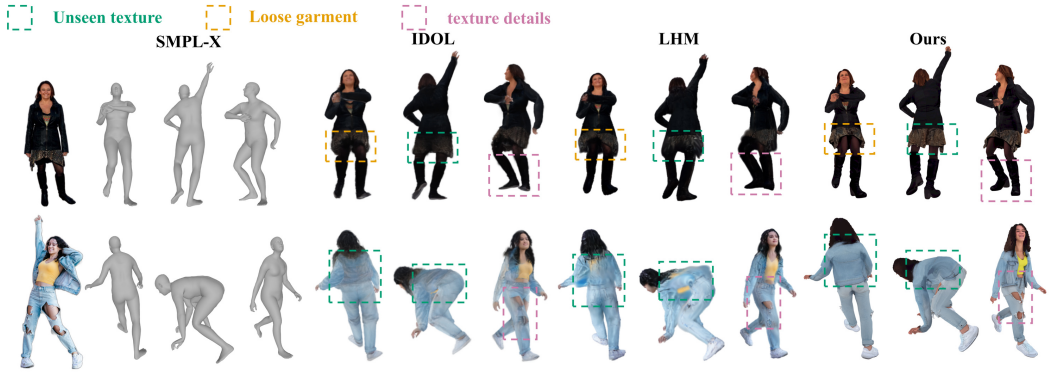

Figure 5: Comparison on Human Animation. NVWA performs better on unseen textures, loose garments, and pose alignment.



Figure 6: Effect of real-scan post-training on generation quality. Post-training improves local geometric fidelity and texture sharpness.

movements. Fig. 5 shows animation results compared to IDOL and LHM. Unlike IDOL and LHM, which require separately estimated SMPL-X parameters, NVWA predicts semantic UV correspondences that directly guide SMPL-X fitting. For each 3D human mesh generated by NVWA, corresponding SMPL-X parameters can be optimized on the basis of UV mesh. As highlighted in Fig. 5, our method better preserves texture details and coherent appearance in initially unobserved regions. Additional qualitative results are provided in Appendix A.5.

User study. We conduct a user study with 27 participants to evaluate perceptual quality. Each participant is shown 15 test cases and asked to compare our method against 5 baselines (LHM, IDOL, PSHuman, TRELLIS2, Pixal3D) across three dimensions: *Overall visual quality.* Users rate which method produces the most realistic and visually pleasing 3D human. *Geometric Accuracy*. Users assess which method best captures the body shape, pose, and fine-grained details (hands, face, clothing). *Image-model Consistency*. Users evaluate the visual consistency between each generated 3D avatar and the input image, focusing on preserving identity and appearance. Tab. 2 confirms that our approach achieves state-of-the-art perceptual quality across multiple evaluation criteria.

5 Ablation Study

Effectiveness of Real-world Scans Post-Training. Training solely on NVWA-100K tends to produce over-smoothed geometry and textures, particularly in fine-scale regions. Following the post-training strategy of LCA (Li et al., 2026b), we therefore introduce an additional refinement stage using real-world 3D scans. As shown in Fig. 6 and Tab. 3, this stage substantially improves geometric fidelity and texture sharpness, demonstrating the importance of real-world scans for recovering high-frequency appearance and geometric details.

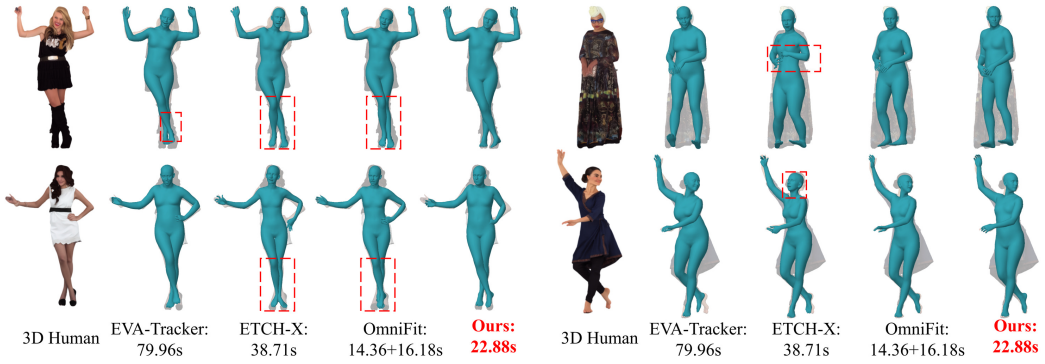

Figure 7: Qualitative comparison of SMPL-X fitting. Geometry-aligned soft rigging enables efficient fitting while maintaining accurate pose alignment.

Table 2: User study on human-generation quality. Values are preference percentages.

Method	Visual quality	Geometric accuracy	Image–model consistency
TRELLIS.2	5.93%	5.93%	5.19%
Pixal3D	4.69%	5.19%	6.42%
IDOL	2.22%	1.23%	0.49%
LHM	6.67%	6.67%	5.68%
PSHuman	18.27%	22.47%	29.63%
Ours	**62.22%**	**58.52%**	**52.59%**

Table 3: Effect of real-scan post-training on X-Humans.

Metric	Without	With
CD ↓	1.75	**1.74**
ULIP-2 ↑	0.3621	**0.3638**
Uni3D ↑	0.3116	**0.3190**
CLIP ↑	0.8613	**0.8749**
LPIPS ↓	0.0819	**0.0813**

Effectiveness and Efficiency of Geometry-Aligned Soft Rigging. A key contribution of NVWA is to decouple expressive free-topology human generation from the topology constraints of a parametric body model, while retaining the semantic correspondence required for animation. To this end, our geometry-aligned soft rigging representation preserves loose garments, hairstyles, and accessories that an SMPL-X-constrained surface cannot faithfully represent. At the same time, it provides reliable correspondences for subsequent SMPL-X fitting, avoiding repeated spatial correspondence searches and reducing ambiguities between geometrically close but semantically distinct body regions. As shown in Fig. 7, our fitting procedure achieves substantially lower runtime than EVA-Tracker (Zhang et al., 2026), ETCH-X (Li et al., 2026c), and the complete OmniFit (Cai et al., 2026) pipeline on NVIDIA RTX 3090, while maintaining accurate pose alignment for challenging poses and loose clothed humans. Consequently, NVWA produces an expressive 3D asset and an animation-enabling soft-rigging representation within a unified generation framework, combining geometric freedom with efficient downstream rigging.

6 Conclusion

We presented NVWA, a native 3D framework that generates animatable human avatars from a single image. Our soft rigging strategy decouples free-topology geometry generation from parametric-body constraints, enabling the model to preserve loose garments, hairstyles, and accessories while retaining full animation capability via efficient SMPL-X fitting and skinning-weight transfer. We also introduced NVWA-100K, a large-scale dataset of explicit 3D humans that adapts generic 3D priors to the clothed-human domain. Experiments demonstrate state-of-the-art geometry, appearance, and animation quality.

Current limitations include dependence on external pose editing and linear blend skinning, which may remain unreliable under extreme occlusion or highly non-rigid garments. Extending the framework toward self-contained pose canonicalization and garment-aware deformation models is a promising direction for future work.

AI USE STATEMENT

We used generative AI tools in both the research workflow and manuscript preparation. OpenAI language-model tools were used to assist with literature discovery and critical feedback on experimental design, draft and revise portions of the manuscript, and improve language clarity. All cited works were subsequently checked by the authors against their original sources, and all technical descriptions, claims, equations, and reported numerical results were manually reviewed and verified.

Generative models were also used as components of our research pipeline. As described in Sec. 3.4, we use GPT-Image2 as an off-the-shelf image editor to transform an input image into an identity- and clothing-preserving A-pose image before 3D generation. The editing prompts and implementation details are provided in the supplementary material. The generated images were manually inspected for identity preservation, garment consistency, pose compliance, and visual artifacts.

Generative AI tools were not used to formulate or prove mathematical claims, fabricate experimental results, or autonomously determine the conclusions of this work. All experiments, quantitative evaluations, and final scientific decisions were conducted or verified by the authors. We take full responsibility for the content of this paper, including all text, claims, code, data, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

Our work follows the ICLR Code of Ethics and aims to advance single-image 3D human generation in a responsible manner. Since NVWA generates animatable 3D avatars from images of real people, the technology may raise concerns regarding privacy, consent, identity impersonation, and the creation of misleading digital content. The generated avatars should not be interpreted as verified representations of an individual's exact body geometry or unobserved appearance, and the method is not intended for biometric identification, surveillance, forensic reconstruction, or deceptive impersonation.

Our training data are constructed from existing image and 3D human datasets, which we use in accordance with their stated licenses and terms of use. We do not collect additional sensitive personal information or attach personally identifying metadata to the constructed 3D assets. The image-space A-pose editing module may synthesize plausible content in regions that are occluded in the input image. Such content represents a model prediction rather than recovery of the person's true hidden appearance. We explicitly communicate this limitation and do not use the synthesized regions as factual or biometric evidence. Our evaluation includes a user study. Participation was voluntary, and participants were informed of the purpose of the study and the use of their responses. We collected only the responses required for aggregate evaluation and did not report personally identifiable participant information. To mitigate potential misuse, released models and data will be accompanied by a responsible-use policy that prohibits non-consensual impersonation, harassment, discriminatory applications, and illegal content. All reported results were reviewed by the authors, who take responsibility for the scientific claims and artifacts presented in this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, data construction pipeline, training procedure, and evaluation protocols in the main paper. The Appendix includes the complete SMPL-X fitting procedure, training configurations and hyperparameters, A-pose editing prompts and settings, and additional fitting and animation results. Upon publication, we will release the source code, configuration files, evaluation scripts, pretrained checkpoints, and the NVWA-100K assets for which redistribution is permitted. These materials will facilitate independent reproduction of the reported generation, fitting, and animation results.

REFERENCES

Rameen Abdal, Wang Yifan, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9441–9451. IEEE, 2024.

Zeyu Cai, Yuliang Xiu, Renke Wang, Zhijing Shao, Xiaoben Li, Siyuan Yu, Chao Xu, Yang Liu, Baigui Sun, Jian Yang, et al. Omnifit: Multi-modal 3d body fitting via scale-agnostic dense landmark prediction. In *European Conference on Computer Vision*, pp. 436–455. Springer, 2026.

Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pp. 557–577. Springer, 2022.

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5795–5805. IEEE, 2021.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16123–16133, 2022.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 119488–119514. Curran Associates, Inc., 2025. doi: 10.52202/085713-3989. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ad444db72317c2a0c50c83e1aa066e27-Paper-Conference.pdf.

Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdiffusion: Volumetric primitives diffusion for 3d human generation. *Advances in Neural Information Processing Systems*, 36:13664–13677, 2023.

Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19925–19936. IEEE, 2023.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14870–14881. IEEE, 2023.

Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12809–12818. IEEE, 2021.

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 *European conference on computer vision*, pp. 1–19. Springer, 2022.

Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12869–12879. IEEE, 2023.

Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21024–21035. IEEE, 2023.

Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM transactions on graphics (TOG)*, 42(4):1–12, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *Advances in neural information processing systems*, 36:10516–10529, 2023.

Dong-Yang Li, Wang Zhao, Yuxin Chen, Wenbo Hu, Meng-Hao Guo, Fang-Lue Zhang, Ying Shan, and Shi-Min Hu. Pixal3d: Pixel-aligned 3d generation from images. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026a.

Junxuan Li, Rawal Khirodkar, Egor Zakharov, Jihyun Lee, Zhaoen Su, Yuan Dong, Julieta Martinez, Kai Li, Qingyang Tan, Takaaki Shiratori, et al. Large-scale codec avatars: The unreasonable effectiveness of large-scale avatar pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18204–18215, 2026b.

Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi, Siyu Xia, Yan-Pei Cao, Wei Xue, et al. Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16008–16018. IEEE, 2025.

Xiaoben Li, Jingyi Wu, Zeyu Cai, Siyuan Yu, Boqian Li, and Yuliang Xiu. Etch-x: Robustify expressive body fitting to clothed humans with composable datasets. In *European Conference on Computer Vision*, pp. 428–446. Springer, 2026c.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6646–6657. IEEE, 2024.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015. doi: 10.1145/2816795.2818013.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9050–9059. IEEE, 2021.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020.

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. 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. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2304–2314, 2019.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 81–90. IEEE, 2020.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16911–16921. IEEE, 2023.

Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1581–1593, 2022.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024a.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *International Conference on Learning Representations*, volume 2024, pp. 33879–33896, 2024b.

Timo Teufel, Pulkit Gera, Xilong Zhou, Umar Iqbal, Pramod Rao, Jan Kautz, Vladislav Golyanik, and Christian Theobalt. Humanolat: A large-scale dataset for full-body human relighting and novel-view synthesis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 29131–29141. IEEE, 2025.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1905–1914. IEEE, 2021.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, yajie bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, and Yao Yao. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 170778–170804. Curran Associates, Inc., 2025. doi: 10.52202/085713-5688. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f9666a092153f281b5116cd1f64b5c91-Paper-Conference.pdf.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14419–14429, 2026.

Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, Shuguang Cui, and Xiaoguang Han. Get3dhuman: Lifting stylegan-human into a 3d generative model using pixel-aligned reconstruction priors. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9253–9263. IEEE, 2023.

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

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13286–13296. IEEE, 2022.

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 512–523. IEEE, 2023.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27081–27091. IEEE, 2024a.

Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human-3diffusion: realistic avatar creation via explicit 3d consistent diffusion models. *Advances in Neural Information Processing Systems*, 37:99601–99645, 2024b.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5742–5752. IEEE, 2021.

Jiawei Zhang, Lei Chu, Jiahao Li, Zhenyu Zang, Chong Li, Xiao Li, Xun Cao, Hao Zhu, and Yan Lu. Bringing your portrait to 3d presence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28468–28480, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Weitian Zhang, Yichao Yan, Yunhui Liu, Xingdong Sheng, and Xiaokang Yang. E 3gen: Efficient, expressive and editable avatars generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6860–6869, 2024.

Xuanmeng Zhang, Jianfeng Zhang, Rohan Chacko, Hongyi Xu, Guoxian Song, Yi Yang, and Jiashi Feng. Getavatar: Generative textured meshes for animatable human avatars. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2273–2282. IEEE, 2023.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6219–6229. IEEE, 2021.

Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 7738–7748. IEEE, 2019.

Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15872–15882. IEEE, 2022.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations*, volume 2024, pp. 46766–46782, 2024a.

Zhenglin Zhou, Fan Ma, Hehe Fan, Zongxin Yang, and Yi Yang. Headstudio: Text to animatable head avatars with 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 145–163. Springer, 2024b.

Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26308–26319. IEEE, 2025.

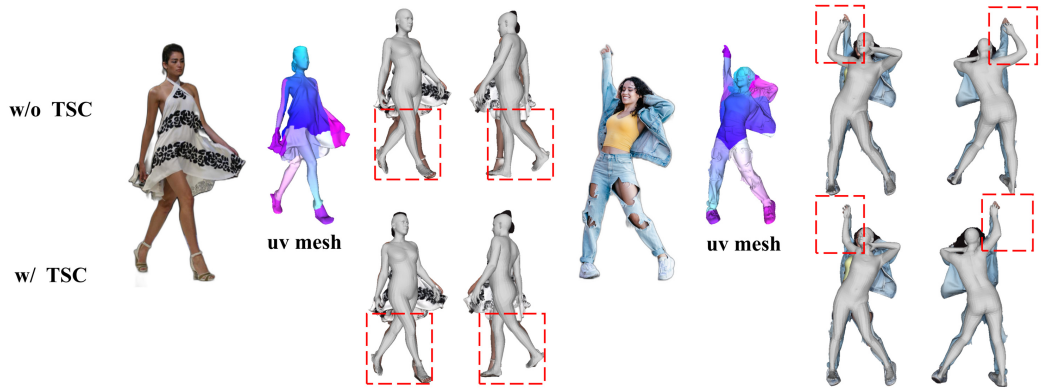

Figure 8: The second-stage Three-View Silhouette Constraint improves SMPL-X pose alignment.

A Appendix

A.1 Details of Geometry-Aligned Soft rigging

Input and notation. This section describes the implementation used to fit SMPL-X to a predicted free-topology human mesh and to transfer the resulting SMPL-X rig to that mesh. Our generation network produces two geometrically corresponding assets: (i) a textured or vertex-colored human mesh used for animation, and (ii) a semantic UV mesh in which the red and green vertex-color channels encode a point in the canonical SMPL-X UV atlas. We represent the predicted semantic UV mesh as

$$M = (X, F, U). \tag{5}$$

Here, X and F define the mesh geometry, and $U = \{\tilde{\mathbf{u}}_i\}$ contains the three-channel soft-rigging attributes. We decode these attributes into 2D semantic UV coordinates $\mathbf{u}_i$ for fitting. For consistency with the main text, we denote the decoded UV coordinates by $\hat{\mathbf{u}}_i^h$ and the predicted vertex positions by $\hat{\mathbf{x}}_i$. The neutral SMPL-X template contains $N_S = 10{,}475$ vertices, canonical UV coordinates $\mathbf{u}_\mathbf{j}^\mathbf{s}$, faces $\mathcal{F}_S$, and linear-blend-skinning weights $\mathbf{W}^S \in \mathbb{R}^{N_S \times 55}$. We optimize the SMPL-X parameters

$$\Theta = \{\boldsymbol{\omega}, \mathbf{t}, s, \boldsymbol{\beta}, \boldsymbol{\phi}_b, \boldsymbol{\phi}_{lh}, \boldsymbol{\phi}_{rh}\}, \tag{6}$$

where $\boldsymbol{\omega} \in \mathbb{R}^3$ is the global orientation, $\mathbf{t} \in \mathbb{R}^3$ is the translation, $s \in \mathbb{R}$ is a global scale, $\boldsymbol{\beta} \in \mathbb{R}^{10}$ is body shape, $\boldsymbol{\phi}_b \in \mathbb{R}^{63}$ is the 21-joint body pose, and $\boldsymbol{\phi}_{lh}, \boldsymbol{\phi}_{rh} \in \mathbb{R}^{45}$ are the two full MANO hand poses. Jaw, eye, and expression parameters are held fixed at zero. The implementation applies translation inside the SMPL-X forward pass and then applies the global scale:

$$\mathbf{V}(\Theta) = s\ \mathrm{SMPLX}(\boldsymbol{\omega}, \mathbf{t}, \boldsymbol{\beta}, \boldsymbol{\phi}_b, \boldsymbol{\phi}_{lh}, \boldsymbol{\phi}_{rh}). \tag{7}$$

Semantic UV correspondence. For fitting, every generated vertex is assigned to its nearest canonical SMPL-X vertex in UV space:

$$a_i = \arg\min_j \|\hat{\mathbf{u}}_i^h - \mathbf{u}_j^s\|_2^2. \tag{8}$$

The lookup is implemented with a two-dimensional k-d tree. Unlike Euclidean nearest-neighbor matching, this semantic correspondence does not exchange vertices between spatially adjacent but anatomically different surfaces, such as an arm and the torso or the two legs. The initial weighted fitting term is

$$\mathcal{L}_{\mathrm{data}}(\Theta) = \frac{1}{Z} \sum_{i=1}^{N} m_i w_i \|\hat{\mathbf{x}}_i - \mathbf{V}_{a_i}(\Theta)\|_2^2, \qquad Z = \sum_i m_i w_i, \tag{9}$$

where $m_i \in \{0, 1\}$ is a robust inlier indicator and w_i is a semantic confidence weight. The default weights are 100 for the face, 5 for hands and feet, and 1 elsewhere. Arm and leg vertices are identified from the summed SMPL-X LBS influence of the corresponding limb joints (threshold

0.25); their default multiplier is 1 but can be increased. UV discontinuities are detected on target-mesh edges whose endpoint UV distance is at least 0.05. The detected set is expanded by two mesh rings and downweighted by 0.25, because a discontinuous semantic code is less reliable near an atlas seam. For fast fitting, we find that the input mesh can contain several hundred thousand vertices, whereas every correspondence terminates at one of only 10,475 SMPL-X vertices. We therefore group all target vertices having the same $a_i = j$. Define

$$W_j = \sum_{i:a_i=j} m_i w_i, \qquad \boldsymbol{\mu}_j = \frac{1}{W_j} \sum_{i:a_i=j} m_i w_i \mathbf{x}_i. \tag{10}$$

Then

$$\sum_i m_i w_i \|\mathbf{x}_i - \mathbf{V}_{a_i}\|_2^2 = \sum_{j:W_j>0} W_j \|\boldsymbol{\mu}_j - \mathbf{V}_j\|_2^2 + C, \tag{11}$$

where C is independent of Θ. Dropping C preserves the objective gradients, minimizer, and Gauss–Newton curvature exactly. Thus the optimizer evaluates at most 10,475 positional constraints without random subsampling.

Two-stage fast fitting. The first fitting phase starts from zero pose, zero translation, unit scale, and zero shape. Its residual vector contains the compressed data residuals and the following soft priors:

$$\mathcal{L}_{\text{fit}} = \mathcal{L}_{\text{data}} + \lambda_\beta \|\boldsymbol{\beta}\|_2^2 + \lambda_s (s-1)^2 + \lambda_{VP} \mathcal{L}_{VP} + \mathcal{L}_{\text{MANO}} + \mathcal{L}_{\text{limits}}. \tag{12}$$

The VPoser term penalizes the difference between the body pose and its VPoser encode–decode reconstruction. The MANO prior penalizes standardized coefficients obtained by projecting the full 45-dimensional pose of each hand onto the MANO PCA basis. Differentiable soft barriers constrain the total finger-joint angle and the twist and total angle of knees, ankles, elbows, and wrists. With robust fitting enabled, λ_β is annealed from 10^{-3} to 10^{-4}, and λ_{VP} from 10^{-1} to 10^{-2}. The MANO-PCA, hand-angle, and body-joint-limit weights are 10^{-3}, 10^{-2}, and 10^{-2}, respectively. We first run an all-correspondence L-BFGS prefit (at most 120 iterations and 180 function evaluations). We then compute the squared fitting error of each input vertex and retain vertices below the 90th percentile. In the current implementation, no semantic region is hard-exempted from this percentile test; the face, hands, and feet are protected through their larger data weights. The compressed constraints are recomputed with the fixed inlier mask, and a robust global L-BFGS solve is run for at most 180 iterations and 260 function evaluations. Strong-Wolfe line search and parameter-dependent step scales are used, and the global scale is clamped to $[0.05, 5]$. In the second stage, the fitting is initialized from the robust solution. We construct front, side, and top orthographic occupancy masks using the (x, y), (z, y), and (x, z) projections of the input vertices. Each mask has resolution 512^2, 8% padding, two-pixel dilation and closing, and hole filling. A Euclidean distance transform gives a differentiable outside-distance field D_k, normalized by the largest body extent. Only fitted SMPL-X vertices outside the observed silhouette are penalized:

$$\mathcal{L}_{\text{sil}}(\Theta) = \frac{\lambda_{sil}}{3N_S} \sum_{k=1}^{3} \sum_{j=1}^{N_S} \left[\max\left(D_k(\pi_k(\mathbf{V}_j)) - \epsilon, 0\right)\right]^2, \tag{13}$$

where π_k is an orthographic projection, $\epsilon = 0.002$, and $\lambda_{sil} = 100$. This one-sided containment objective allows clothing to remain outside the body while discouraging the fitted body from protruding beyond the reconstructed surface. All 170 fitting variables remain optimizable in this phase. To prevent the silhouette term from destroying the semantic solution, we add quadratic anchors to the first-phase global orientation, body pose, and both hand poses:

$$\begin{aligned}\mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{data}} + \mathcal{L}_{\text{sil}} + \lambda_\omega \|\boldsymbol{\omega} - \boldsymbol{\omega}^{(1)}\|_2^2 \\ + \lambda_b \|\boldsymbol{\phi}_b - \boldsymbol{\phi}_b^{(1)}\|_2^2 + \lambda_h \sum_{q \in \{lh, rh\}} \|\boldsymbol{\phi}_q - \boldsymbol{\phi}_q^{(1)}\|_2^2 + \lambda_{\beta,2} \|\boldsymbol{\beta}\|_2^2.\end{aligned} \tag{14}$$

We use $\lambda_\omega = \lambda_b = \lambda_h = 0.02$ and $\lambda_{\beta,2} = 3 \times 10^{-4}$. Translation and scale have no second-phase anchor. Optimization uses L-BFGS with strong-Wolfe line search, up to 120 iterations, 180 function evaluations, and history size 50. As shown in Fig. 8, the second-stage three-view silhouette constraint improves the alignment of the legs under loose clothing and the arms under complex articulation, resulting in more accurate SMPL-X fitting.

Skinning-weight transfer. Fitting uses a nearest UV vertex, but rig transfer uses the continuous SMPL-X UV triangulation. For each target UV point $\mathbf{u}_i$, a k-d tree over SMPL-X UV-triangle centroids retrieves eight candidate triangles. We compute the exact closest point on every candidate triangle and select the closest triangle $f_i = (a_i, b_i, c_i)$ with barycentric coordinates $\boldsymbol{\alpha}_i$. The initial target skinning weights are

$$\mathbf{w}_i^0 = \alpha_{i1}\mathbf{W}_{a_i}^S + \alpha_{i2}\mathbf{W}_{b_i}^S + \alpha_{i3}\mathbf{W}_{c_i}^S, \qquad \sum_{j=1}^{55} w_{ij}^0 = 1. \tag{15}$$

This interpolation avoids piecewise-constant weights and preserves smooth transitions across the canonical body. We compute the 3D distance between each target vertex and its corresponding point on the fitted SMPL-X surface. Vertices with distances below the 90th percentile are treated as reliable anchors. This 3D validation is important because a plausible UV code can still be incorrect at loose clothing, self-contact, or a discontinuity. Let $\mathcal{A}$ be the reliable anchor vertices. Unreliable weights are filled by minimizing a regularized biharmonic energy with Dirichlet boundary conditions:

$$\min_{\mathbf{W}} \operatorname{tr}(\mathbf{W}^\mathsf{T}\mathbf{Q}\mathbf{W}), \qquad \mathbf{W}_{\mathcal{A}} = \mathbf{W}_{\mathcal{A}}^0, \tag{16}$$

where

$$\mathbf{Q} = -\mathbf{L} + \mathbf{L}\mathbf{M}^{-1}\mathbf{L} + 10^{-6}\mathbf{I}, \tag{17}$$

$\mathbf{L}$ is twice the cotangent Laplacian and $\mathbf{M}$ is the Voronoi mass matrix. A nearest-reliable-vertex fallback is used if the sparse solve is ill-conditioned. We then perform ten local smoothing iterations with blend factor 0.2 over the inpainted area and nearby vertices within 5% of the mesh bounding-box diagonal. The default pipeline additionally applies five whole-mesh adjacency-smoothing iterations with blend factor 0.5. Negative or invalid weights are removed and every weight vector is renormalized to sum to one.

A.2 Details of training

For each of the four DiT models, we insert LoRA adapters into the attention projections $(to_{qkv}, to_q, to_{kv}, to_{out})$ and the two linear projections of each MLP block ($mlp.0$ and $mlp.2$). We use a rank of $r = 64$, a scaling factor of $\alpha = 128$, and a LoRA dropout rate of 0.05. We additionally train and save $AdaLN_modulation.1$, while keeping all other pretrained DiT parameters frozen.

During post-training, we perform mixed-data training on 5K real-world human scans from THuman2.1, Twindom, and CustomHumans (Yu et al., 2021; Ho et al., 2023), together with 2K synthetic assets randomly selected from NVWA-100K, which reduces the synthetic-to-real domain gap and yields better-defined facial features and garment folds, together with sharper texture details. We only train the geometry generative model and texture generative model at a learning rate of 5×10^{-5}.

A.3 Details of Image-Space A-Pose Editing

Meshes generated from self-contacting poses may contain spurious connections between nearby body and garment surfaces. These incorrect connections are severely stretched during large articulations, causing body–garment adhesion, distorted limbs, and surface tearing. Post-processing based on stretched-edge removal and Poisson reconstruction cannot reliably recover the missing geometry or correct the ambiguous topology and may cause artifacts. Our Image-space A-pose Editing stage transforms the input into an identity-preserving, open-limb pose before 3D generation, reducing self-contact and plausibly completing occluded regions. As shown in Fig. 9, this strategy produces cleaner surfaces and substantially more stable, higher-quality animation under large pose changes. We use the same prompt for all test images. The prompt specifies the target pose and explicitly constrains identity, body proportions, clothing, framing, and image appearance. The exact prompt is reproduced below.

```
Perform a strict identity-preserving pose edit of the input image.

CHANGE ONLY THE PERSON'S BODY POSE:

Repose the person into a symmetric, front-facing, standard A-pose.
```

```
- The head, face, gaze, chest, torso, and pelvis face directly toward
    the camera.
- Keep the head upright, with the original facial expression.
- Keep the shoulders level and symmetric.
- Extend both arms straight outward and downward from the shoulders.
- Abduct each arm approximately 35-45 degrees away from the side of the
    torso, so the arms form the two downward-sloping sides of a capital
    letter A.
- Keep both hands below elbow level and well below shoulder level.
- Do not make either arm horizontal or raise the arms into a T-pose.
- Keep both elbows fully extended, wrists neutral, and palms facing
    downward.
- Keep both hands and all fingers natural, complete, separated, and
    fully visible.
- Keep both legs straight.
- Abduct each leg outward approximately 20 degrees from the body's
    vertical centerline, forming an approximately 40-degree angle
    between the two legs.
- Keep both knees straight.
- Place both feet on the same ground plane, pointing forward or only
    slightly outward.
- Keep a clear visible gap between each arm and the torso.
- Keep a clear visible gap between the two legs.
- There must be no self-contact or overlap between the arms, torso,
    hands, or legs.
- Keep the complete head, hands, fingers, legs, and feet inside the
    image.
- Center the full body with sufficient empty margin around the entire
    silhouette.

PRESERVE EVERYTHING ELSE EXACTLY:

Preserve the same person's exact identity and facial likeness,
    including facial geometry, eyes, eyebrows, nose, lips, ears, skin
    tone, age, expression, makeup, and all visible facial details.

Preserve the exact hairstyle, hairline, hair color, and hair
    accessories.

Preserve the original body identity and proportions, including shoulder
    width, torso length, waist, limb length, and limb thickness.

Preserve the exact original clothing and shoes: garment type, cut,
    length, fabric, colors, patterns, printed graphics, logos,
    embroidery, seams, buttons, accessories, and material appearance.
    Only adjust the wrinkles and folds that are physically required by
    the new pose. Do not redesign, replace, simplify, recolor, lengthen,
    or remove any garment.

Preserve the original camera viewpoint, perspective, lighting, shadows,
    background, color balance, photographic style, and image quality.

DO NOT:

- change the face, identity, body shape, clothing, hairstyle, or
    background;
- add or remove clothing, accessories, jewelry, text, or objects;
- create extra, missing, duplicated, fused, or deformed limbs or
    fingers;
- create fabric bridges or webbing between the arms and torso;
- fuse the two legs or connect the clothing across the leg gap;
- crop the head, hands, fingers, legs, or feet;
- output a skeleton, collage, comparison image, caption, border, or
    watermark.
```

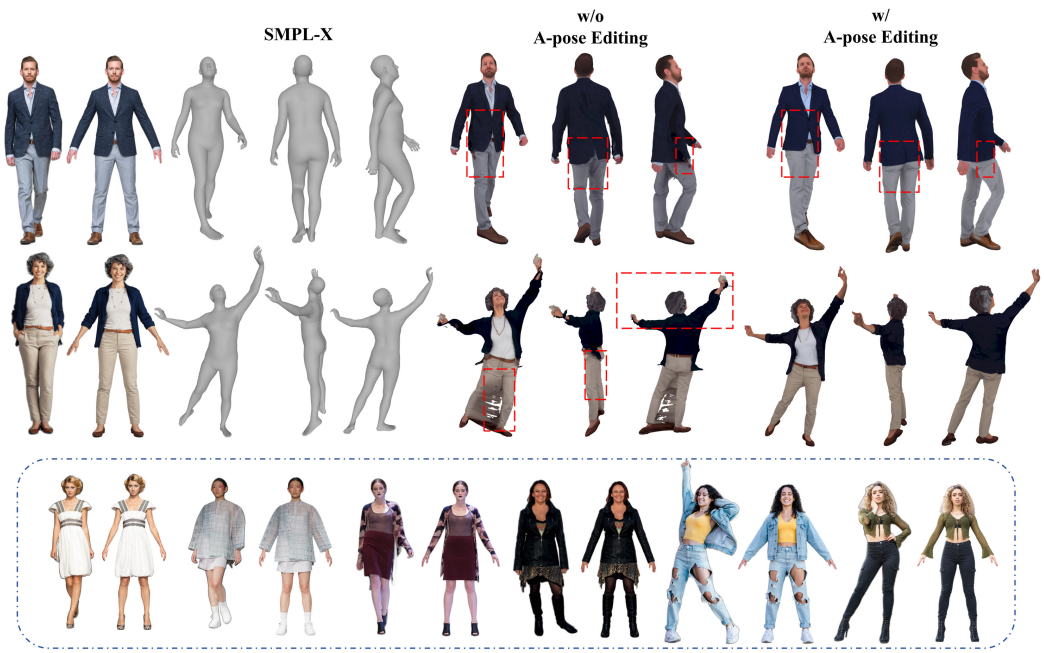

Figure 9: Image-Space A-Pose Editing reduces body–garment adhesion and deformation artifacts. The bottom row shows the 2D editing results.

```
The output must look like the same original photograph with only the
    person's body pose changed.
```

A.4 Details of Evaluation Metrics

We employ a suite of complementary metrics to evaluate semantic consistency, native 3D alignment, geometric fidelity, and texture reconstruction quality. CLIP measures the visual-semantic agreement between the input image and multi-view renderings of the generated asset. ULIP-2 and Uni3D evaluate image-to-3D consistency directly in a joint image/point-cloud embedding space. For X-Humans, where ground-truth scans are available, we additionally report Chamfer Distance (CD) on aligned visible outer surfaces and LPIPS on aligned textured RGB renderings. For the in-the-wild benchmark, which has no ground-truth 3D scans or novel-view images, LPIPS is instead computed only between the input image and the frontal rendering of the generated asset. All methods within a comparison are evaluated on exactly the same sample set. Details of the metrics are provided below.

CLIP Score. The CLIP score measures the semantic and appearance consistency between the conditioning image and the generated 3D asset. For each sample, we use the 36 predefined uniformly spaced orbit views produced by each method. Both the conditioning image and rendered views are processed by the same OpenCLIP image encoder. Let f_I denote the L2-normalized embedding of conditioning image I, and let f_{R_v} denote the normalized embedding of view R_v. The per-sample score is

$$S_{\text{CLIP}}(I, \mathcal{R}) = \frac{1}{36} \sum_{v=0}^{35} f_I^\top f_{R_v}. \tag{18}$$

We average the per-sample scores to obtain the result for each method. This evaluation does not use a ground-truth mesh or ground-truth rendering, and it does not perform per-sample orientation search. A higher CLIP score indicates better consistency between the input image and the appearance of the generated asset across viewpoints.

ULIP-2 and Uni3D Scores. ULIP-2 and Uni3D evaluate the agreement between the conditioning image and the generated asset from a native 3D perspective. We first convert each colored mesh into a 10,000-point xyzRGB point cloud. Specifically, we sample 100,000 candidate points uniformly

with respect to triangle surface area and then apply farthest-point sampling to retain 10,000 spatially distributed points. The RGB value of each point is obtained by querying the corresponding base-color texture using its barycentrically interpolated UV coordinate. If the asset stores baked vertex colors instead of a texture map, RGB is obtained by barycentric interpolation of the triangle vertex colors. For ULIP-2, the point cloud is encoded by the official ULIP-2 Colored PointBERT model trained for 10k xyzRGB inputs, while the conditioning image is encoded using its paired OpenCLIP `ViT-bigG-14` image tower. For Uni3D, we use the official Uni3D-G point-cloud encoder and the paired `EVA02-E-14-plus` OpenCLIP image tower. In both cases, the point-cloud and image embeddings are L2-normalized, and the score is their cosine similarity. Scores are averaged over all evaluated samples. Higher ULIP-2 and Uni3D scores indicate stronger image-to-3D semantic and appearance alignment. These two metrics require an explicitly colored mesh; methods represented only by 3D Gaussian primitives without a recoverable colored surface mesh are therefore excluded from these comparisons.

Chamfer Distance. We use Chamfer Distance to evaluate geometric fidelity on X-Humans, for which ground-truth scans and fitted SMPL-X parameters are available. Before computing the distance, we explicitly resolve coordinate-system, scale, and orientation differences using ICP. Following registration, we evaluate only the visible outer shell rather than directly sampling all mesh triangles. The metric is computed in the independently normalized unit-cube coordinate system and therefore evaluates normalized shape fidelity rather than absolute body scale. We report $\mathrm{CD} \times 10^4$ for readability. A lower value indicates better geometric agreement with the ground-truth scan.

LPIPS. We use the Learned Perceptual Image Patch Similarity (LPIPS) metric to evaluate textured RGB reconstruction quality under two dataset-specific protocols. Unlike pixel-wise metrics, LPIPS measures the distance between deep feature representations and is therefore less dominated by minor low-level photometric differences, such as small variations in color, illumination, and local texture. To enable a fairer comparison across generative models, we establish corresponding views and normalize rendering conditions whenever ground-truth geometry is available. A lower LPIPS value indicates greater perceptual similarity.

X-Humans. For methods that provide textured meshes, we apply the same per-sample ground-truth and prediction transformations estimated during the Chamfer Distance evaluation. The aligned prediction and textured ground-truth scan are then rendered from the same 36 viewpoints used for CLIP evaluation using identical cameras, lighting conditions, resolutions, and white backgrounds.

In-the-wild. Because in-the-wild inputs do not provide ground-truth 3D scans or ground-truth novel-view images, we evaluate only reconstruction of the observed frontal view. The reported score for each method is the mean of its per-sample frontal-view LPIPS distances.

A.5 More Qualitative Results

Figs. 10-14 present additional generation, animation and fitting results on diverse in-the-wild images covering a wide range of identities, body shapes, poses, hairstyles, garments, and accessories. IDOL-Apose and LHM-Apose mean we introduce the pose-edited A-pose image as their input. NVWA consistently generates high-fidelity textured human meshes while preserving the identity, appearance, and fine-grained clothing details of the input images. In particular, its free-topology representation faithfully captures loose garments and structures that deviate substantially from the underlying body surface. Moreover, the proposed geometry-aligned soft rigging representation enables reliable SMPL-X fitting and semantically consistent skinning-weight transfer. The resulting avatars can be driven by diverse target motions while maintaining coherent geometry and appearance, demonstrating the generalization and animation robustness of our method.

A.6 More samples of NVWA-100K

Figs. 15- 16 present additional representative samples from NVWA-100K. Each identity is visualized from both frontal and rear viewpoints. The dataset covers diverse body shapes, poses, hairstyles, and clothing styles, including casual wear, formal wear, sportswear, uniforms, long dresses, coats, layered garments, and accessories such as hats and backpacks. The corresponding front and rear renderings further demonstrate the multi-view appearance consistency of the textured explicit as-



Figure 10: More generation results from single real-world human Images.

sets. These diverse 3D humans provide direct geometry and appearance supervision for adapting a native 3D generative prior to the clothed-human domain.



Figure 11: More animation results compared with IDOL and LHM.



Figure 12: More animation results from single real-world human Images.



Figure 13: More animation results from single real-world human Images.



Figure 14: More fitting results from single real-world human Images.



Figure 15: More samples in NVWA-100K.



Figure 16: More samples in NVWA-100K.