DIRECT REWARD FINE-TUNING ON POSES FOR SINGLE IMAGE TO 3D HUMAN IN THE WILD

Anonymous authors
Paper under double-blind review

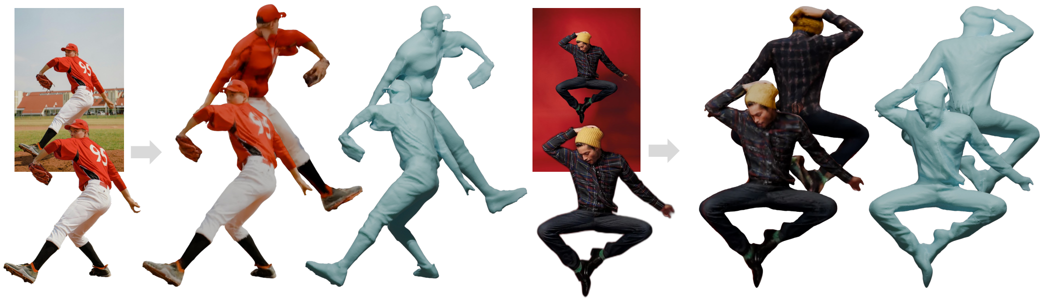

Figure 1: We propose DRPOSE, a method to post-train a multi-view diffusion model for enhanced posture of reconstructed 3D humans in dynamic and acrobatic scenarios.

ABSTRACT

Single-view 3D human reconstruction has achieved remarkable progress through the adoption of multi-view diffusion models, yet the recovered 3D humans often exhibit unnatural poses. This phenomenon becomes pronounced when reconstructing 3D humans with dynamic or challenging poses, which we attribute to the limited scale of available 3D human datasets with diverse poses. To address this limitation, we introduce DRPOSE, a Direct Reward fine-tuning algorithm on Poses, which enables post-training of a multi-view diffusion model on diverse poses without requiring expensive 3D human assets. DRPOSE trains a model using only human poses paired with single-view images, employing a direct reward fine-tuning to maximize POSESCORE, which is our proposed differentiable reward that quantifies consistency between a generated multi-view latent image and a ground-truth human pose. This optimization is conducted on DRPOSE15K, a novel dataset that was constructed from an existing human motion dataset and a pose-conditioned video generative model. Constructed from abundant human pose sequence data, DRPOSE15K exhibits a broader pose distribution compared to existing 3D human datasets. We validate our approach through evaluation on conventional benchmark datasets, in-the-wild images, and a newly constructed benchmark, with a particular focus on assessing performance on challenging human poses. Our results demonstrate consistent qualitative and quantitative improvements across all benchmarks.

1 INTRODUCTION

3D human models are essential assets across multiple industries, including visual media production (such as games and movies), product and industrial design, and e-commerce platforms for fashion. While multi-view scanning systems and manual design processes currently dominate 3D human crafting workflows, single-view 3D human reconstruction technology has garnered attention due to rapid technical advances and its practical advantages in scenarios where capturing multiple camera angles is either impractical or impossible.

Recent advances in this technology have been driven by the adoption of image-to-multi-view (I2MV) diffusion models, which have enhanced reconstruction quality for occluded body parts invisible in the input image (Pan et al., 2024; Peng et al., 2024; Li et al., 2024b; He et al., 2024; Xue et al., 2024; Ho et al., 2023a). This approach typically employs a two-stage pipeline: first generating multi-view images from a single input using a diffusion model, then lifting these views into 3D space through either implicit reconstruction (Saito et al., 2019; Ho et al., 2023a) or explicit reconstruction techniques (Li et al., 2024b; Palfinger, 2022; Xiu et al., 2022a). Compared to previous works, which directly reconstruct a 3D structure from the input-view feature (Saito et al., 2019; 2020) or works utilizing an estimated SMPL model (Xiu et al., 2022b;a), multi-view diffusion-based approaches have the benefit of using more fine-detailed cues for the unseen parts from the input-view.

Despite these advancements, a bottleneck persists that limits real-world applicability. Reconstructed 3D humans often exhibit unnatural postures, especially when target poses are dynamic and challenging, such as extreme athletic movements or acrobatic postures. We argue that this limitation stems from the limited scale of publicly available training datasets (Yu et al., 2021; Han et al., 2023; Ho et al., 2023b) with diverse poses. This scarcity arises from the costs of recruiting diverse subjects and capturing them in varied poses using multi-view stereo setups, which are further compounded by privacy concerns that complicate the release of public data.

Our key insight to overcome this challenge is that, instead of requiring expensive 3D human assets for training, we can leverage available 3D pose sequence data (Lin et al., 2023) and a pose-conditioned video generative model (Men et al., 2025) to construct a DRPOSE15K, a dataset consisting of single-view images for input and corresponding ground-truth poses. To this end, we introduce DRPOSE, a method to post-train an I2MV model on this dataset using a *direct reward fine-tuning algorithm* (Liu et al., 2024; Clark et al., 2023; Xu et al., 2023; Prabhudesai et al., 2024). In DRPOSE, given an input image, a pre-trained I2MV model generates multi-view latent images through an iterative denoising process. Then, the latents are compared with the ground-truth 3D pose to compute POSESCORE, our proposed differentiable reward function that quantifies the consistency between them. The pretrained I2MV model is optimized to maximize POSESCORE, across the DRPOSE15K, which has a broader pose distribution coverage than the existing 3D human datasets.

Our evaluation demonstrates that I2MV models fine-tuned with DRPOSE achieve improvements in single-view 3D human reconstruction quality both quantitatively and qualitatively. These improvements are consistent across all datasets, including conventional benchmarks (Yu et al., 2021; Ho et al., 2023b), in-the-wild images, and MIXAMORP, our new evaluation benchmark designed to assess performance on complex and dynamic human poses.

Our key contributions are:

- We propose DRPOSE, a novel post-training approach for enhancing the alignment of an image-to-multi-view (I2MV) model with natural poses in dynamic and complex scenarios.
- We construct DRPOSE15K, a dataset comprising human poses from a motion dataset (Lin et al., 2023) paired with generated single-view images conditioned on each pose.
- Through quantitative evaluation, we demonstrate that our method achieves consistent improvements across all datasets, including conventional benchmarks and our proposed MIXAMORP.

2 RELATED WORKS

2.1 SINGLE-VIEW 3D HUMAN RECONSTRUCTION

Single-view 3D human reconstruction remains a long-standing challenge in computer vision and graphics. Early approaches focused on recovering parametric human models (Loper et al., 2023; Pavlakos et al., 2019) but often lacked fine-grained details such as clothing and facial features (Bogo et al., 2016; Zhang et al., 2021; 2023a; Sun et al., 2021). A major advance was introduced by PIFu (Saito et al., 2019), which demonstrated that detailed 3D human shapes could be learned from a single image using implicit functions trained on 3D scan datasets. This inspired numerous extensions, including methods that (1) utilize normal maps to enhance surface quality (Saito et al., 2020; Xiu et al., 2022b;a), (2) utilizing SMPL prior (Xiu et al., 2022b;a; Zhang et al., 2023b; Zhuang et al., 2025), (3) recover relightable textures (Alldieck et al., 2022), and (4) generate animation-ready avatars (Huang et al., 2020; He et al., 2021; Peng et al., 2024). Recently, generative models have further advanced the field by improving reconstruction quality for previously unseen views by adopting *score distillation sampling* (Huang et al., 2023; Wang et al., 2025; AlBahar et al., 2023; Wang et al., 2024) or training a multi-view diffusion model (Pan et al., 2024; Peng et al., 2024; Li et al., 2024b; He et al., 2024; Xue et al., 2024; Hu et al., 2025). However, when these models receive images with out-of-distribution poses as input, they show results that exhibit unnatural postures. To address this, we propose a new approach that leverages motion data (Lin et al., 2023) to augment pose coverage and fine-tune multi-view diffusion models, thereby improving performance on diverse poses.

2.2 Direct Reward Fine-tuning of Diffusion Model

Recent research has explored methods for post-training diffusion models to align them with human preferences better, building on the success of reinforcement learning techniques in large language models. This alignment process typically involves three key components: (1) starting with a pretrained text-to-image diffusion model, (2) developing a reward model that evaluates attributes such as aesthetic quality, detail fidelity, and semantic alignment, and (3) optimizing the diffusion model to maximize these reward signals. Initial approaches utilized reinforcement learning (RL) objectives to maximize human preferences, though these methods are non-differentiable (Lee et al., 2023; Black et al., 2023; Fan et al., 2023). Building on human preference data, researchers have developed differentiable neural networks that can evaluate input images (Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023). Leveraging these advances, direct reward fine-tuning methods have recently emerged that post-train diffusion models using differentiable reward scores (Prabhudesai et al., 2024; Clark et al., 2023; Wu et al., 2024), demonstrating faster convergence compared to RL-based approaches. In this work, we adopt DRTune (Wu et al., 2024), a state-of-the-art reward fine-tuning method, as the foundation for DrPose.

3 Preliminaries

3.1 Image-to-multi-view (I2MV) Diffusion Model

We adopt an image-to-multi-view (I2MV) diffusion model to our single-view 3D human reconstruction pipeline to provide fine-detailed cues for the unseen regions of the human subject from the input view. Era3D (Li et al., 2024a), a state-of-the-art I2MV model, introduces a row-wise attention layer as an additional layer to the stable diffusion's denoising U-Net. This layer performs self-attention across pixels in the same row, spanning all multi-view images, thereby maintaining multi-view consistency during generation. Unlike previous multi-view attention layers (Shi et al., 2023; Wang & Shi, 2023; Höllein et al., 2024) that apply self-attention across all pixels in the multi-view images, the row-wise approach reduces computational overhead from $O(N^2S^4)$ to $O(N^2S^3)$, where S denotes the spatial resolution and N represents the number of views. For our base I2MV diffusion model, we adopt the denoising U-Net from PSHuman (Li et al., 2024b), which extends Era3D (Li et al., 2024a) by incorporating a body-face cross-scale diffusion architecture that enhances the quality of face region generation.

3.2 3D Human Reconstruction with Explicit Carving

As illustrated in Figure 2, we employ an explicit carving in our pipeline to reconstruct 3D humans from multi-view images generated by our post-trained diffusion model, following Li et al. (2024b). The pipeline generates both normal maps and RGB images across multiple viewpoints using a diffusion model conditioned on the input view. 3D human mesh recovery then proceeds through three sequential steps: SMPL-X initialization, differentiable remeshing (Palfinger, 2022), and appearance fusion. This approach delivers superior geometric detail compared to methods using pretrained implicit networks (Ho et al., 2023a; Pan et al., 2024)

4 Method

This section describes our proposed method for aligning an image-to-multi-view (I2MV) diffusion model to natural postures in dynamic or complex cases, thereby enhancing the quality of its integrated single-view 3D human reconstruction pipeline. We begin in Sec. 4.1 by describing the

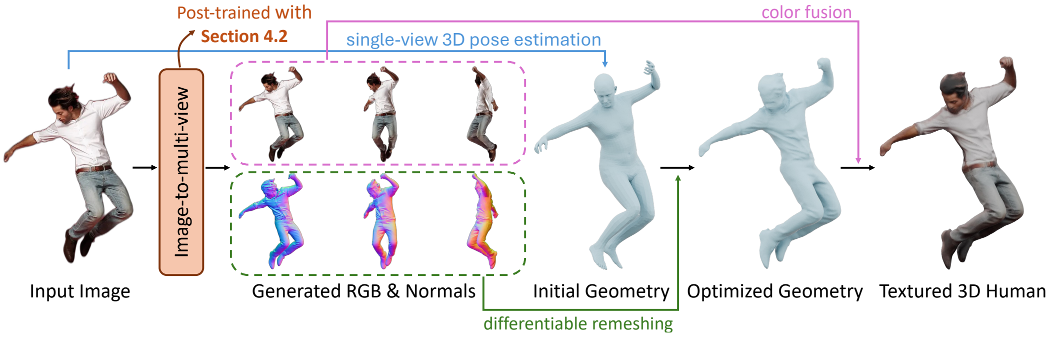

Figure 2: Overview of our 3D human reconstruction pipeline. In this pipeline, the multi-view normal and RGB images are generated from the input image using a image-to-multi-view (I2MV) diffusion model. Then these images are converted into 3D representation using explicit human carving (Li et al., 2024b). In this work, we propose post-training the I2MV diffusion model to achieve better alignment with accurate poses in dynamic and acrobatic scenarios. For clarity, only 3 of the 6 multi-view images are displayed for normal maps and RGB images.

construction of DrPose15K, our proposed training dataset with diverse pose coverage. Sec. 4.2 then presents DrPose, which enables post-training of an I2MV model on DrPose15K.

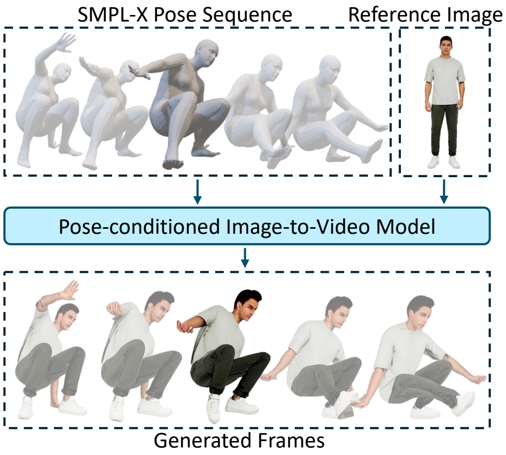

Figure 3: Construction process for DrPose15K. We employ a pose-conditioned image-to-video model Men et al. (2025) to generate input-view images corresponding to the ground-truth poses.

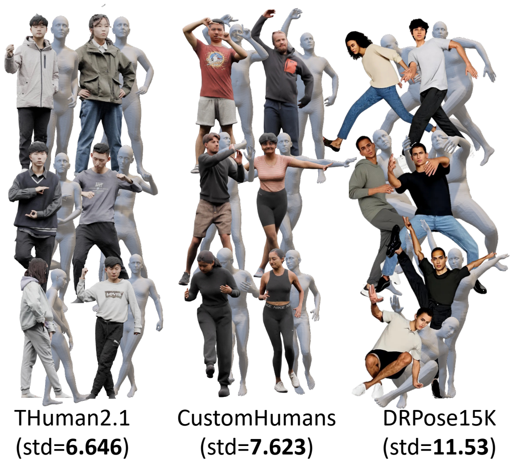

Figure 4: Comparison of pose diversity between conventional 3D human datasets (Yu et al., 2021; Ho et al., 2023b) and our proposed DrPose15K. Our dataset has a higher standard deviation of SMPL-X joint locations than other datasets.

4.1 Construction of DrPose15K

We construct DrPose15K, a training dataset containing dynamic and challenging 3D human poses paired with single-view images, by leveraging Motion-X (Lin et al., 2023), a human motion dataset and MIMO (Men et al., 2025), a pose-conditioned image-to-video(I2V) model as illustrated in Figure 3. From the Motion-X dataset, we utilize the AIST (Li et al., 2021) subset due to its comprehensive coverage of diverse pose distributions. To reduce redundancy from the 300K available poses, we apply farthest point sampling to select 1.5K poses. Then, we add the 9 temporal neighbors for each selected pose to create a pose sequence for input to the MIMO, yielding a total of 15K poses. Finally, we use MIMO to animate full-body human images from Photos (2025) according to these pose sequences, generating corresponding single-view images for each 3D pose in our dataset.

To quantitatively assess the pose diversity of DrPose15K compared to conventional 3D human datasets (Ho et al., 2023b; Yu et al., 2021), we compute the standard deviation of SMPL-X joint

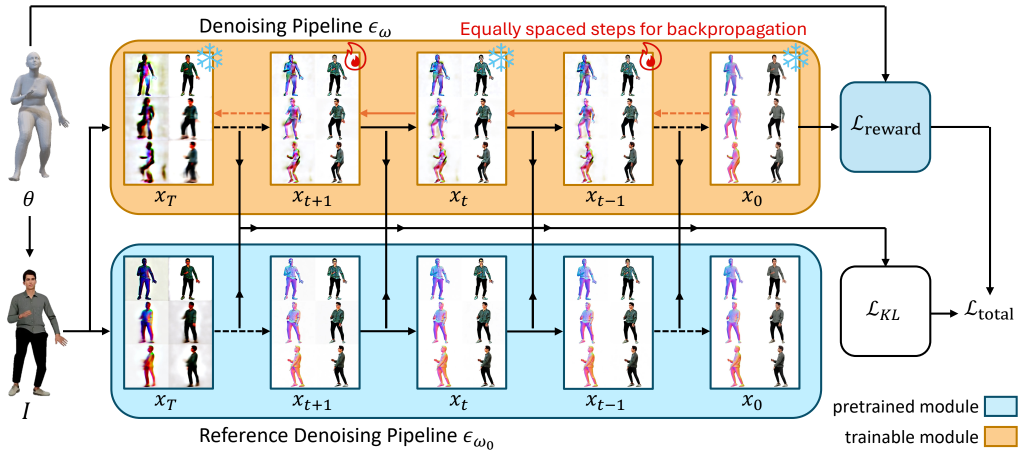
Denoising Pipeline ϵ_ω
Equally spaced steps for backpropagation
θ
I
x_T
x_{t+1}
x_t
x_{t-1}
x_0
$\mathcal{L}_{\text{reward}}$
$\mathcal{L}_{KL}$
$\mathcal{L}_{\text{total}}$
pretrained module
trainable module
Reference Denoising Pipeline ϵ_{ω_0}

Figure 5: Overview of DRPOSE. Given a 3D human pose θ and input image I (generated from θ as described in Sec 4.1), the denoising multi-view U-Net ϵ_ω is trained to minimize $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reward}} + w_{\text{KL}} \cdot \mathcal{L}_{\text{KL}}$. Here, $\mathcal{L}_{\text{reward}}$ measures the distance between θ and the generated latent image x_0, while $\mathcal{L}_{\text{KL}}$ computes the KL divergence between ϵ_w and the frozen initial U-Net ϵ_{w_0} (Sec 4.2). For clarity, only 3 of 6 multi-view images are shown for normal maps and RGB visualization.

positions across each dataset, focusing exclusively on the 22 body joints while excluding facial and hand joints. Note that for conventional datasets, we include both training and test splits in this analysis. As shown in Figure 4, DRPOSE15K exhibits a **1.73×** larger standard deviation compared to THuman2.1 (Yu et al., 2021). Moreover, with 14.7K poses compared to 647 in CustomHumans and 2,445 in THuman2.1, TrainSet provides broader pose distribution coverage.

4.2 DRPOSE (DIRECT REWARD FINE-TUNING ON POSES)

We introduce DRPOSE, an algorithm to post-train an I2MV diffusion model on DRPOSE15K, denoted as $D = \{I_i, \theta_i\}$, where I_i, θ_i are an input image and the ground-truth human pose. The core idea is to maximize POSESCORE, our proposed differentiable reward that quantifies consistency between the generated multi-view latent images from I_i and θ_i, better aligning the pretrained I2MV diffusion model to diverse poses in D.

DRPOSE builds upon previous direct reward fine-tuning algorithms (Wu et al., 2024; Prabhudesai et al., 2024). The method generates latent images $\boldsymbol{x}_0$ at timestep $t = 0$ through an iterative denoising process, then computes the reward loss L_{reward} using POSESCORE, a differentiable reward function denoted as r. Since maintaining gradients for all timesteps would require prohibitive GPU memory, we sample a subset of timesteps t_{train} for gradient computation. Following DRTune (Wu et al., 2024), DRPOSE samples equally spaced timesteps from the full denoising trajectory, enabling optimization of early denoising steps while maintaining computational efficiency.

Algorithm 1 DRPOSE

```
Dataset: Image-pose pairs $D = \{I_i, \theta_i\}$
Inputs: I2MV diffusion model with initial weights $\omega_0$, reward model $r$, the number of training timesteps $K$, maximum early stop timestep $m$
Initialize $\omega = \omega_0$
while not converged do
    $s = \text{randint}(1, T - K\lfloor \frac{T}{K} \rfloor)$
    $t_{\text{train}} = \{s + i\lfloor \frac{T}{K} \rfloor \mid i = 0, 1, \ldots, K-1\}$
    $t_{\text{min}} = \text{randint}(1, m)$
    $(I, \theta) \sim D$
    $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$
    $\mathcal{L}_{\text{KL}} = 0$
    for $t = T, \cdots, 1$ do
        $\hat{\epsilon} = \epsilon_\omega(\texttt{stop\_grad}(\boldsymbol{x}_t), I, t)$
        if $t \notin t_{train}$ then
            $\hat{\epsilon} = \texttt{stop\_grad}(\hat{\epsilon})$
        else
            $\hat{\epsilon}_0 = \epsilon_{\omega_0}(\texttt{stop\_grad}(\boldsymbol{x}_t), I, t)$
            $\mathcal{L}_{\text{KL}} = \mathcal{L}_{\text{KL}} + \mathbb{E}(||\hat{\epsilon} - \hat{\epsilon}_0||)$
        $\hat{\boldsymbol{x}}_0 = (\mathbf{x}_t - \sigma_t \hat{\epsilon})/\alpha_t$
        if $t == t_{min}$ then
            break
        $\boldsymbol{x}_{t-1} = \alpha_{t-1}\hat{\boldsymbol{x}}_0 + \sigma_{t-1}\hat{\epsilon}$
    $\mathcal{L}_{\text{reward}} = 1 - r(\hat{\boldsymbol{x}}_0, \theta)$
    $\omega \leftarrow \omega - \eta \nabla_\omega (\mathcal{L}_{\text{reward}} + w_{\text{KL}} \cdot \mathcal{L}_{\text{KL}})$
```

To address the reward hacking problem, where reward scores increase during training while image quality degrades, DRPOSE incorporates a KL divergence regularization term L_{KL} in addition to L_{reward}. This regularization computes $\mathbb{E}(||\hat{\epsilon} - \hat{\epsilon}_0||)$, where $\hat{\epsilon}$ represents the predicted noise from the trainable diffusion model at some timestep $t \in t_{\text{train}}$, and $\hat{\epsilon}_0$ is the corresponding prediction from the initial diffusion model. This constraint prevents the model's generated images from deviating excessively from its original results while optimizing for reward maximization.

To summarize, DRPOSE operates the steps in Algorithm 1 to minimize the following objective:

$$\min_{\omega} \mathbb{E}_{(I,\theta)\sim D} \left[\mathcal{L}_{\text{reward}}\left(I, \theta\right) + w_{\text{KL}} \cdot \mathcal{L}_{\text{KL}}\left(I\right)\right]. \tag{1}$$

Differentiable Reward. To quantify the consistency a multi-view latent image $\boldsymbol{x}_0$ and a GT pose θ, we develop POSESCORE, a differentiable reward model denoted as r. To compute the consistency, it first projects both $\boldsymbol{x}_0$ and θ to the $\hat{I}_{\text{skel}}$ and I_{skel}, images where the human skeletal structure are drawn. To convert $\boldsymbol{x}_0$ into $\hat{I}_{\text{skel}}$, a U-Net based skeletal image predictor g_{skel} is pretrained on the existing 3D human datasets (Ho et al., 2023b; Yu et al., 2021). Moreover, I_{skel} can be drawn from θ, by drawing the projected the 3D human joints $J(\theta)$ from the pose parameter θ into the image planes same with the generated images' viewpoints. Then the reward is compute as follows:

$$r(\boldsymbol{x}_0, \theta) = -\mathbb{E}(||\hat{I}_{\text{skel}} - I_{\text{skel}}||) = -\mathbb{E}(||g_{\text{skel}}(\boldsymbol{x}_0) - \mathcal{R}(J(\theta))||), \tag{2}$$

where $\mathcal{R}$ is the rendering of the 3D human joints into the skeletal images into the viewpoints of $\boldsymbol{x}_0$.

5 EXPERIMENTS

5.1 IMPLEMENTATION DETAILS

Denoising U-Net We initialize our model ϵ_{ω_0} with the denoising U-Net architecture from PSHuman (Li et al., 2024b). The model is fine-tuned on four NVIDIA H100 GPUs using a batch size of 4 with gradient accumulation over 2 steps for 5.5K iterations. During training, we employ the DDIM sampler with $T = 20$ total denoising steps and $K = 2$ training steps. We set the maximum early stop timestep to $m = 8$ and weight the KL divergence loss as $w_{\text{KL}} = 0.01$. For computing $\mathcal{L}_{\text{KL}}$, we use mean squared error to estimate $||\hat{\epsilon} - \hat{\epsilon}_0||$. At inference time, we use the DDIM sampler with $T = 40$ denoising steps and apply classifier-free guidance (Ho & Salimans, 2022) with a scale of 3.0.

Differentiable Reward For computing the reward, we use binary cross entropy loss and LPIPS to estimate $||\hat{I}_{\text{skel}} - I_{\text{skel}}||$. The skeletal images $\hat{I}_{\text{skel}}$ and I_{skel} both have 23 channels, with each channel corresponding to one skeleton. We use THuman2.1 (Yu et al., 2021) and the training subset of CustomHumans (Ho et al., 2023b) as our training datasets, comprising approximately 3K scans. To get six-view normal and color images, we render the 3D scans using Blender's Cycles engine (Community, 2018) with an orthographic camera configuration. The reward model is trained on four NVIDIA RTX 6000 Ada GPUs with a batch size of 16 over 10K iterations.

5.2 SINGLE-VIEW 3D HUMAN RECONSTRUCTION

Baselines & Benchmarks We compare our approach against single-view 3D human reconstruction methods guided by SMPL (Xiu et al., 2022a; Ho et al., 2023a), as well as multi-view diffusion-based methods (Wu et al., 2023; Li et al., 2024a;b).

- **ECON** (Xiu et al., 2022a) estimates front and back depth maps using an estimated SMPL-X prior, then fuses these depth maps for a complete 3D human body. It does not support texture reconstruction and trains its depth estimation network on 500 scans from THuman2.0 (Yu et al., 2021). The depth estimation network is trained on 500 scans from THuman2.0.
- **SiTH** (Ho et al., 2023a) generates 512×512 px. RGB images for front and back views using an estimated SMPL-X prior, subsequently converting them to 3D via an SDF network. The diffusion model is trained on THuman2.0.
- **Human3Diffusion** (Xue et al., 2024) produces four 256×256 px. RGB multi-view images, which are then converted to 3D using a 3DGS reconstruction network. The multi-view

Table 1: Quantitative comparisons of geometry quality on single-view human reconstruction benchmarks. Our proposed benchmark MixamoRP is described in Appendix A.2. Era3D* represents the original Era3D model fine-tuned on CustomHumans and THuman2.1 training splits using conventional DDPM loss. Ours (Era3D) denotes the Era3D model post-trained with our proposed DrPose on DrPose15K.

	THuman2.1-test			CustomHumans-test			MixamoRP		
Method	CD↓	NC↑	f-Score↑	CD↓	NC↑	f-Score↑	CD↓	NC↑	f-Score↑
ECON	57.8809	0.6760	13.5307	70.0954	0.6552	10.4112	187.5267	0.5655	4.7752
SiTH	64.8460	0.6677	14.2759	77.5391	0.6504	11.5578	146.5484	0.5764	6.8088
Era3D*	54.2934	0.7018	15.1518	62.3912	0.7056	14.0601	111.0537	0.6163	8.6145
PSHuman	48.0357	0.7202	17.8297	57.0701	0.7099	15.4065	101.8600	0.6244	9.5673
Ours (Era3D)	39.8191	0.7387	19.3195	43.1307	0.7425	18.9756	90.8153	0.6307	10.3593
Ours	37.6248	0.7434	20.7005	44.7405	0.7381	18.1897	94.3054	0.6274	9.8742

Table 2: Quantitative evaluation of 3D human reconstruction quality. Six RGB views evenly distributed in azimuth are rendered to compute appearance metrics. Our proposed benchmark MixamoRP is described in Appendix A.2. Era3D* represents the original Era3D model fine-tuned on CustomHumans and THuman2.1 training splits using conventional DDPM loss. Ours (Era3D) denotes the Era3D model post-trained with our proposed DrPose on DrPose15K.

	THuman2.1-test			CustomHumans-test			MixamoRP		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SiTH	16.8538	0.7884	0.1743	15.7267	0.7773	0.2098	13.5855	0.7604	0.2748
Era3D*	18.7502	0.8226	0.1380	18.9253	0.8355	0.1326	17.5337	0.8623	0.1519
PSHuman	19.0605	0.8259	0.1285	19.0814	0.8373	0.1273	17.6624	0.8641	0.1497
Ours (Era3D)	19.1135	0.8406	0.1242	19.1135	0.8406	0.1242	17.5568	0.8662	0.1475
Ours	19.3110	0.8303	0.1243	19.3404	0.8411	0.1224	17.6631	0.8646	0.1471

diffusion model is trained on 6K human scans combining public datasets (Yu et al., 2021; Ho et al., 2023b; Han et al., 2023) and commercial datasets (AXYZ design, 2023; Renderpeople, 2023; Treedy, 2023; Twindom, 2023).

- **Era3D** (Li et al., 2024a) generates six 512×512 px. normal and RGB images using a diffusion network trained on Objaverse (Deitke et al., 2023). For fair comparison, we fine-tune this model on 3K scans from THuman2.1 and CustomHumans (Ho et al., 2023b) datasets.
- **PSHuman** (Li et al., 2024b) produces six 768×768 px. normal and RGB images using a diffusion network trained on THuman2.1 and CustomHumans datasets.

All models above are evaluated quantitatively in the following three benchmarks:

- **THumans2.1-test** contains 60 human scans selected from the full THumans2.1 (Yu et al., 2021) dataset. The split follows Li et al. (2024b).
- **CustomHumans-test** contains 60 human scans selected from the full CustomHumans dataset, which consists of 600 human scans. The split follows Ho et al. (2023a).
- **MixamoRP** is our proposed benchmark containing 60 human scans, constructed by assigning 60 distinct poses collected from Mixamo animation, to 15 different Renderpeople 3D models, with 4 poses per a model(see Appendix A.2 for more details).

Test splits from CustomHumans (Ho et al., 2023b) and THuman2.1 (Yu et al., 2021) are commonly used benchmarks for evaluating single-view 3D human reconstruction methods. While these benchmarks include dynamic poses such as dancing or jumping, they lack extremely complex poses (see Figure 4) like breakdancing or bat swinging. To establish new evaluation criteria for 3D human reconstruction under extreme pose variations, we introduce MixamoRP, a novel benchmark specifically designed to assess reconstruction performance on challenging pose configurations. See

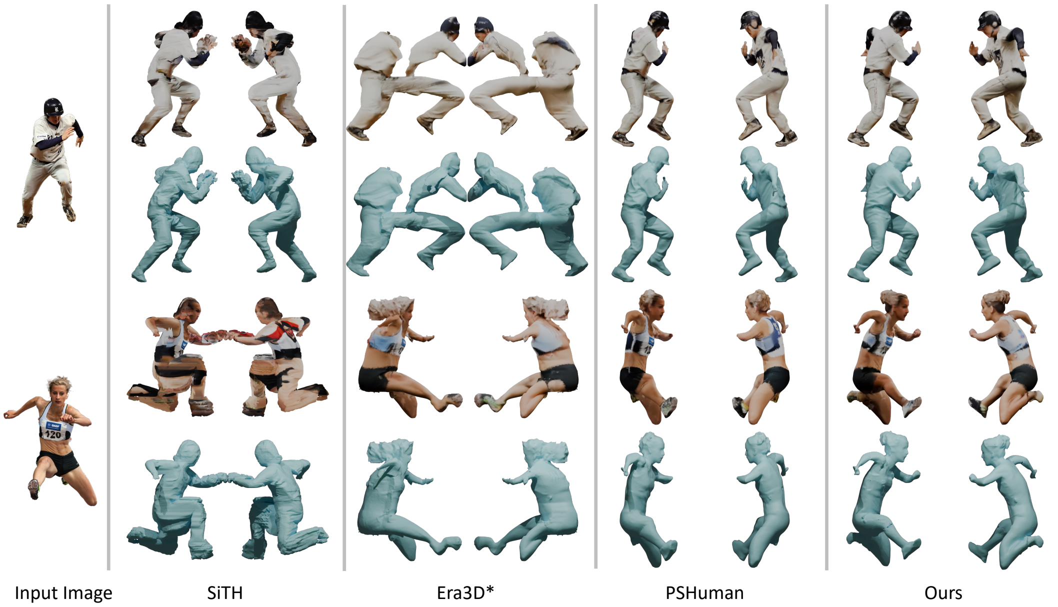

Figure 6: Qualitative evaluation on the internet-source images. Era3D* denotes Era3D fine-tuned on CustomHumans and THuman2.1 datasets.

Appendix A.2 for the complete list of Renderpeople (Renderpeople, 2023) characters and corresponding Mixamo (Inc., 2025) animations used in our dataset.

Evaluation protocol For each mesh scan, we render input images from 3 evenly distributed azimuthal views, yielding 180 input views per benchmark. To evaluate geometric accuracy, we report three metrics in Table 1: Chamfer Distance (CD), Normal Consistency (NC), and F-Score. For computing Chamfer Distance, we uniformly sample 100K points per mesh.

For appearance evaluation, we report three metrics in Table 2: PSNR, SSIM, and LPIPS. To compute these metrics, we render images of both the prediction and ground truth from 6 evenly distributed azimuthal views that are distinct from the input views.

Results As Table 1 and Table 2 presents, our results demonstrate that DrPose consistently improves reconstruction quality of the base model across all benchmarks. This is thanks to our proposed DrPose's ability to enhance the accuracy of reconstructed posture on diverse poses, as seen in the Figure 6 and Figure 7.

Ablation Study on the base model We conduct an ablation study on the base model by post-training Era3D* using DrPose. As reported in Table 1 and Table 2, the Era3D-based model shows similar performance across all benchmarks. However, since the PSHuman-based model shows better results on face regions qualitatively, we chose PSHuman as our base model.

5.3 Analysis on PoseScore

Table 3: **Quantitative evaluation of g_{skel} in PoseScore**

Benchmark	PSNR	SSIM	LPIPS
THuman2.1-test	22.4807	0.9337	0.0580
CustomHumans-test	24.4081	0.9536	0.0430

In the Figure 8, we provide the analysis of the trained g_{skel} of PoseScore introduced in Section 4.2. Figure 8 and Table 3. The evaluation is conducted on the SMPL and scan mesh pairs on the test splits of CustomHumans and THuman2.1. The scan meshes are rendered into the multi-view normal and color images to be converted into the latent images via PSHuman's VAE. These latent images are fed into g_{skel}, producing the skeletal images. The metrics and qualitative results show that g_{skel} reliable enough to use it as a measure for the consistency between latent images and poses.

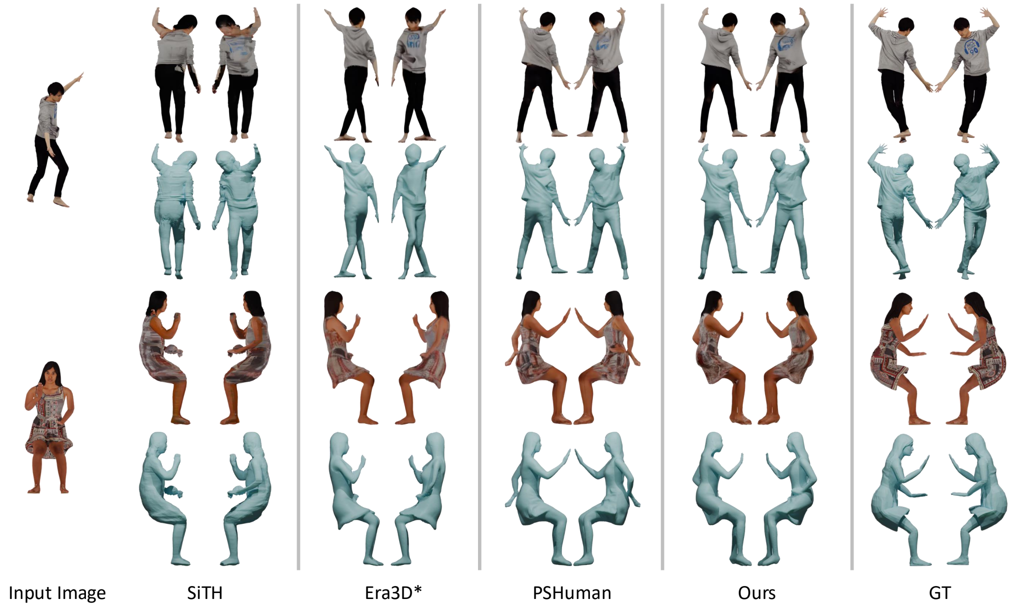

Figure 7: Qualitative evaluation on the CustomHumans dataset. Era3D* denotes Era3D fine-tuned on CustomHumans and THuman2.1 datasets.

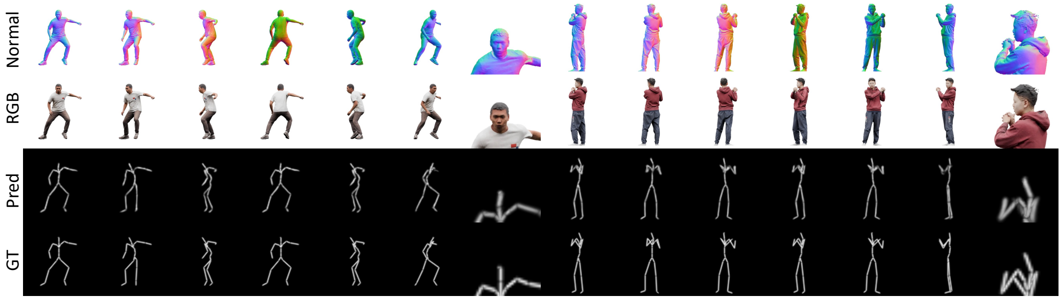

Figure 8: Visualization of g_{skel} in POSESCORE. g_{skel} converts the multi-view latent images encoded from the normal and RGB images using the base model's VAE.

6 CONCLUSION

We propose a novel approach to improve the pose accuracy of 3D humans reconstructed by multi-view diffusion models. Our method comprises three key contributions: (1) DRPOSE15K, a dataset featuring diverse poses with corresponding single-view images, (2) DRPOSE, an algorithm that enables post-training of multi-view diffusion models on this dataset; and (3) MIXAMORP, a benchmark for evaluating reconstruction under challenging poses. Our post-trained model shows consistent quality improvements across all benchmarks.

Limitations Similar to previous single-image-to-3D human modeling approaches, our pipeline requires segmented input images. When input images contain imperfect segmentation, artifacts such as floating geometry appear in the boundary regions of the generated 3D humans, as illustrated in Figure 9.

Although DRPOSE employs gradient stopping and gradient checkpointing techniques, it requires substantial GPU memory, as it generates 24 images of size 768×768 px, through an iterative denoising process to compute POSESCORE. We believe improved efficiency in future multi-view diffusion models will alleviate this issue.

Ethics Statement

Demographic Bias Our base model, PSHuman (Li et al., 2024b), is trained on THuman2.1 (Yu et al., 2021) and CustomHumans (Ho et al., 2023b), which exhibit demographic imbalances. THuman2.1 contains 2,445 human subjects who are predominantly of Asian ethnicity, while CustomHumans, though more ethnically diverse, comprises only 647 subjects. This imbalanced representation may result in biased reconstruction performance that favors demographics overrepresented in the training data, leading to reduced quality and accuracy for underrepresented groups.

Potential for Misuse The generated 3D human models pose risks for creating misleading or harmful content. These reconstructions can be integrated into 3D scenes and animated using standard rigging techniques, potentially enabling the creation of for disinformation or deepfake content.

Industrial Impact The automation capabilities of image-to-3D human modeling technology may impact employment in creative industries, affecting 3D artists, character designers, and digital content creators who specialize in human modeling While this technology can enhance productivity and accessibility, it also raises questions about the displacement of skilled professionals.

Reproducibility Statement

DrPose15K is constructed from the publicly available Motion-X dataset (Lin et al., 2023) and MIMO model (Men et al., 2025). MixamoRP is constructed from scans of RenderPeople (Renderpeople, 2023) and motions from Mixamo (Inc., 2025); while both resources are available, RenderPeople is a commercial product. In Section 4, we explain the high-level concepts underlying our approach and provide pseudocode and experimental details in Algorithm 1 and Section 5.1 to ensure reproducibility.

References

Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *CVPR*, 2022.

AXYZ design. Axyz, November 2023. URL `https://secure.axyz-design.com`. Accessed on 7, 34, 36.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. *arXiv.org*, 2023. doi: 10.48550/arxiv.2305.13301.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 561–578. Springer, 2016.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv.org*, 2023. doi: 10.48550/arxiv.2309.17400.

Blender Online Community. Blender - a 3d modelling and rendering package, 2018. URL `http://www.blender.org`.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, M. Ryu, Craig Boutilier, P. Abbeel, M. Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *arXiv.org*, 2023. doi: 10.48550/arxiv.2305.16381.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2023)*, June 2023.

Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. *IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/iccv48922.2021.01086.

Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Hao-Wen Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement. *arXiv.org*, 2024. doi: 10.48550/arxiv.2408.14211.

Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. *arXiv.org*, 2023a. doi: 10.48550/arxiv.2311.15855.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5043–5052, 2024.

Shoukang Hu, Takuya Narihira, Kazumi Fukuda, Ryosuke Sawata, Takashi Shibuya, and Yuki Mitsufuji. Humangif: Single-view human diffusion with generative prior. *arXiv preprint arXiv:2502.12080*, 2025.

Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. *arXiv.org*, 2023. doi: 10.48550/arxiv.2308.08545.

Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. *Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/cvpr42600.2020.00316.

Adobe Inc. Mixamo, 2025. URL `https://www.mixamo.com/`. Accessed on September 24, 2025. Provides online tools for animating 3D characters, automatic character rigging, and motion captured animations for games, films, and interactive experiences.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv.org*, 2023. doi: 10.48550/arxiv.2305.01569.

Kimin Lee, Hao Liu, M. Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, P. Abbeel, M. Ghavamzadeh, and S. Gu. Aligning text-to-image models using human feedback. *arXiv.org*, 2023. doi: 10.48550/arxiv.2302.12192.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, Wenping Wang, Qi-fei Liu, and Yi-Ting Guo. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv.org*, 2024a. doi: 10.48550/arxiv.2405.11616.

Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. Pshuman: Photorealistic single-view human reconstruction using cross-scale diffusion. *arXiv preprint arXiv:2409.10141*, 2024b.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13401–13412, 2021.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.

Buhua Liu, Shitong Shao, Bao Li, Lichen Bai, Zhiqiang Xu, Haoyi Xiong, James Kwok, Sumi Helal, and Zeke Xie. Alignment of diffusion models: Fundamentals, challenges, and future. *arXiv preprint arXiv:2409.07253*, 2024.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866, 2023.

Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21181–21191, 2025.

Werner Palfinger. Continuous remeshing for inverse rendering. *Computer Animation and Virtual Worlds*, 33(5):e2101, 2022. doi: https://doi.org/10.1002/cav.2101. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cav.2101.

Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. *arXiv.org*, 2024. doi: 10.48550/arxiv.2406.12459.

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.

H. Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. *ACM transactions on graphics*, 2024. doi: 10.1145/3658217.

Generated Photos. Generated photos datasets. https://generated.photos/datasets, 2025. Free to use for academic research.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. In *ICLR*, 2024.

Renderpeople. Renderpeople, November 2023. URL https://renderpeople.com/. Accessed on 7, 34.

Shunsuke Saito, Zeng Huang, Ryota Natsume, S. Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *IEEE International Conference on Computer Vision*, 2019. doi: 10.1109/iccv.2019.00239.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. *computer vision and pattern recognition*, 2020. doi: 10.1109/cvpr42600.2020.00016.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11179–11188, 2021.

Treedy. Treedy, November 2023. URL https://treedys.com/. Accessed on 7, 34.

Twindom. Twindom, November 2023. URL `https://web.twindom.com/`. Accessed on 7, 10, 34, 36.

Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv.org*, 2023. doi: 10.48550/arxiv.2312.02201.

Wentao Wang, Hang Ye, Fangzhou Hong, Xue Yang, Jianfu Zhang, Yizhou Wang, Ziwei Liu, and Liang Pan. Geneman: Generalizable single-image 3d human reconstruction from multi-source human data. *arXiv preprint arXiv:2411.18624*, 2024.

Zilong Wang, Zhiyang Dou, Yuan Liu, Cheng Lin, Xiao Dong, Yunhui Guo, Chenxu Zhang, Xin Li, Wenping Wang, and Xiaohu Guo. Wonderhuman: Hallucinating unseen parts in dynamic 3d human reconstruction. *arXiv preprint arXiv:2502.01045*, 2025.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. *IEEE International Conference on Computer Vision*, 2023. doi: 10.1109/iccv51070.2023.00200.

Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 108–124. Springer, 2024.

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. Econ: Explicit clothed humans optimized via normal integration. *Computer Vision and Pattern Recognition*, 2022a. doi: 10.1109/cvpr52729.2023.00057.

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, 2022b.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv.org*, 2023. doi: 10.48550/arxiv.2304.05977.

Yuxuan Xue, Xianghui Xie, R. Marin, and Gerard Pons-Moll. Human 3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. *arXiv.org*, 2024. doi: 10.48550/arxiv.2406.08475.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5746–5756, 2021.

Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11446–11456, 2021.

Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023a.

Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. *arXiv.org*, 2023b. doi: 10.48550/arxiv.2312.06704.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26308–26319, 2025.

A Appendix

A.1 Additional Results

We provide additional qualitative comparisons in Figure 13 and 14.

A.2 MixamoRP

As mentioned in Section 5.2, we constructed MixamoRP using Renderpeople's rigged 3D models and Mixamo animations. Table 5 details character names, animations, descriptions, and frame indices for reproducibility. Each row represents a unique mesh with challenging poses. Figure 15 visualizes the dataset.

A.3 Limitations

Our pipeline inherits limitations from prior single-image-to-3D approaches, shown in Figure 9. First, imperfect input segmentation causes floating geometry artifacts at boundaries. Second, while improving overall shape and pose, our method struggles with fine details like hands.

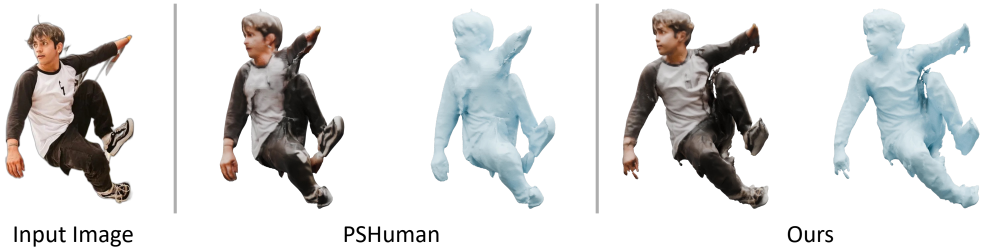

Figure 9: Our pipeline is sensitive to the quality of the segmented masks, producing artifacts.

A.4 Analysis on Efficiency

We report the latency of each model for reconstructing a single sample in Table 4. Ours have the same efficiency as PSHuman since we use it as our base model, while Ours (Era3D) have the same efficiency as the Era3D.

Table 4: Reconstruction latency per sample for each model.

ECON	SiTH	Era3D	PSHuman	Ours (Era3D)	Ours
183.27 sec.	117.35 sec.	15.24 sec.	42.70 sec.	15.24 sec.	42.70 sec.

A.5 Network Architecture

Figures 10, 11, and 12 show the network design for the Denoising U-Net in the multi-view diffusion used in the pipeline illustrated in Fig 2, the Denoising U-Net in the MIMO Men et al. (2025), a pose-conditioned video generator, and the skeletal image predictor in the PoseScore introduced in Sec 4.2.

A.6 Use of Large Language Models

We employ a large language model to refine the manuscript text by correcting grammatical errors and enhancing sentence fluency. The LLM is not involved in research ideation, methodology development, experimental design, or the generation of original content. All intellectual contributions, including the research direction, analyses, and conclusions, are made entirely by the authors.

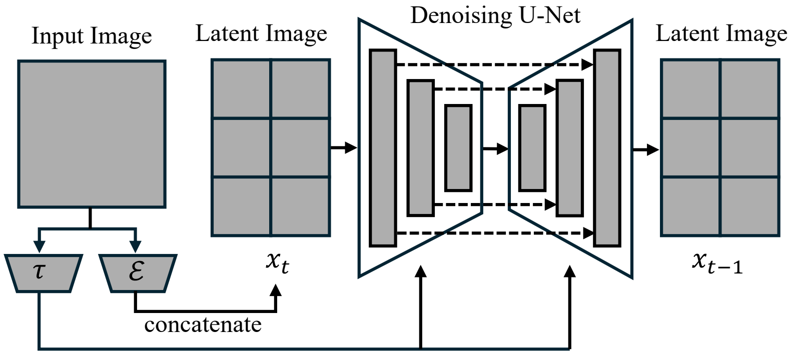

Figure 10: Architecture of the Denoising U-Net for multi-view diffusion in our pipeline inspired from Li et al. (2024b) (illustrated in Fig 2). The denoising U-Net follows the architecture of PSHuman (Li et al., 2024b). The input image is conditioned into the denoising process through two parallel pathways: (1) A VAE encoder $\mathcal{E}$ encodes the input image, which is then concatenated with the latent image x_t. (2) A CLIP image encoder τ encodes the input image, and the generated tokens are fed into the cross-attention layers of the denoising U-Net.

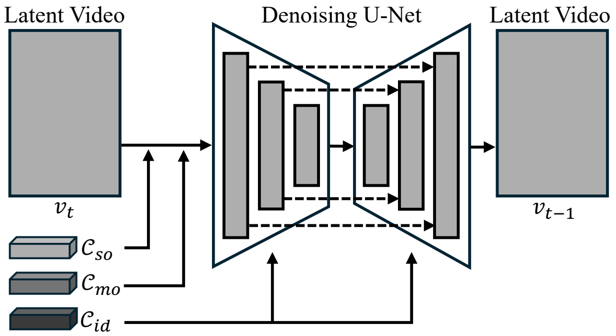

Figure 11: Architecture of the denoising U-Net for the pose-conditioned image-to-video model used in the DrPose15K construction process (illustrated in Fig. 3). The denoising U-Net follows the architecture of MIMO (Men et al., 2025) and takes three conditioning signals: (1) scene code $\mathcal{C}_{\text{so}}$, (2) motion code $\mathcal{C}_{\text{mo}}$, and (3) identity code $\mathcal{C}_{\text{id}}$. The scene code $\mathcal{C}_{\text{so}}$ is first concatenated with the latent video v_t and then added to the motion code $\mathcal{C}_{\text{mo}}$. The identity code $\mathcal{C}_{\text{id}}$ is fed into the cross-attention layers of the denoising U-Net. Note that the temporal layers of the denoising U-Net (Guo et al., 2023) are omitted in this figure.

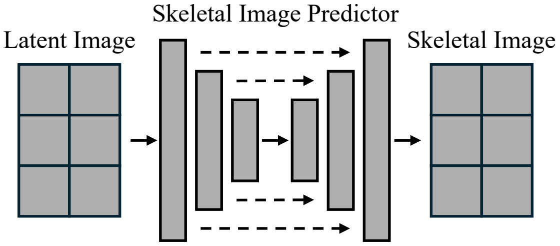

Figure 12: Architecture of the Skeletal Image Predictor in our PoseScore introduced in Sec 4.2. The network follows a U-Net architecture with an encoder-decoder structure. Multi-view latents are processed through initial convolutions, then flattened and passed through four downsampling blocks (reducing spatial resolution from 64×64 to 4×4 while increasing channels from 32 to 512), followed by four upsampling blocks with skip connections that restore the original resolution. The output produces predicted skeletal images for all views simultaneously.

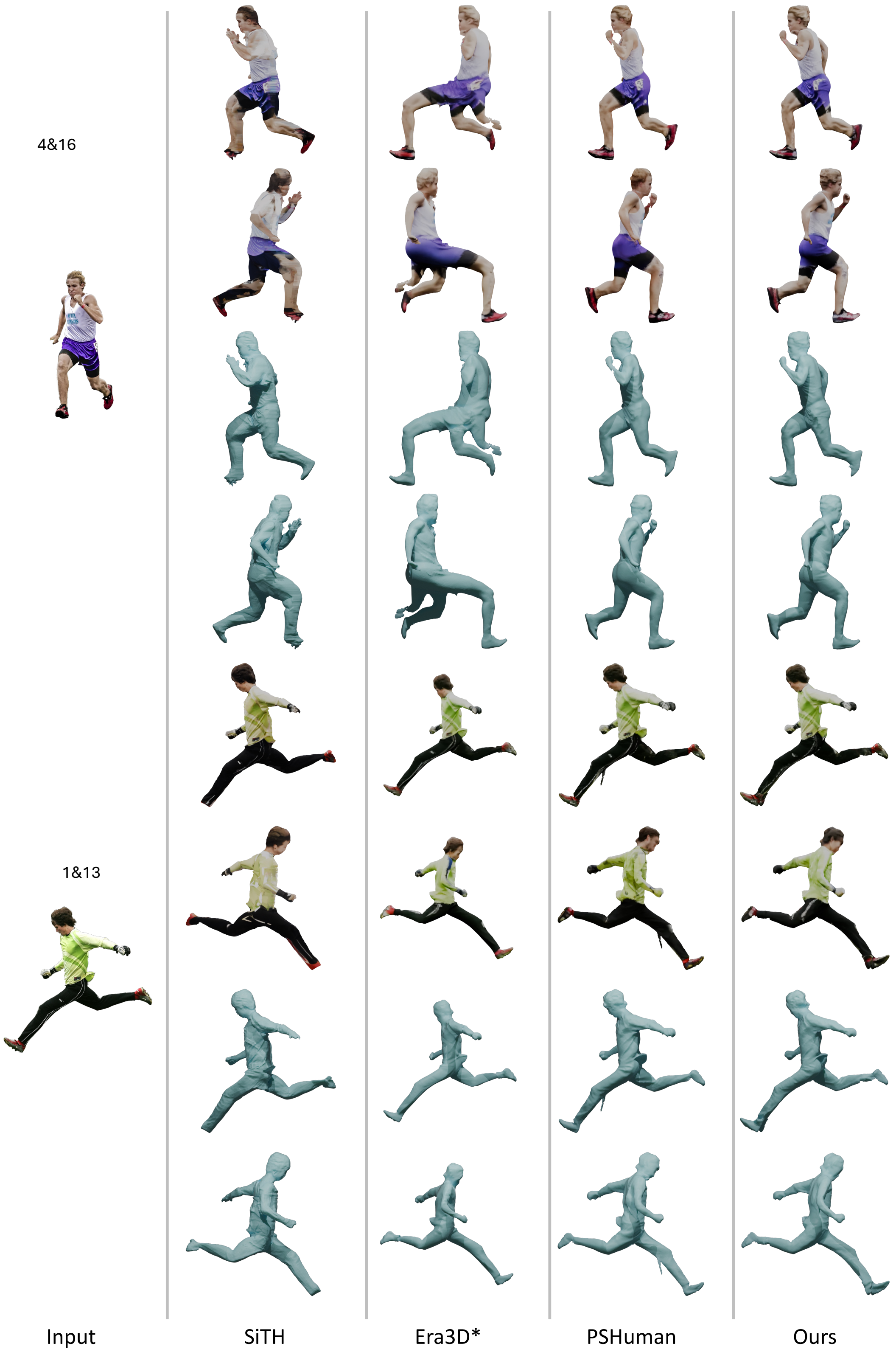

Figure 13: Additinoal qualitative evaluation on the internet-source images. Era3D* denotes Era3D fine-tuned on CustomHumans and THuman2.1 datasets.

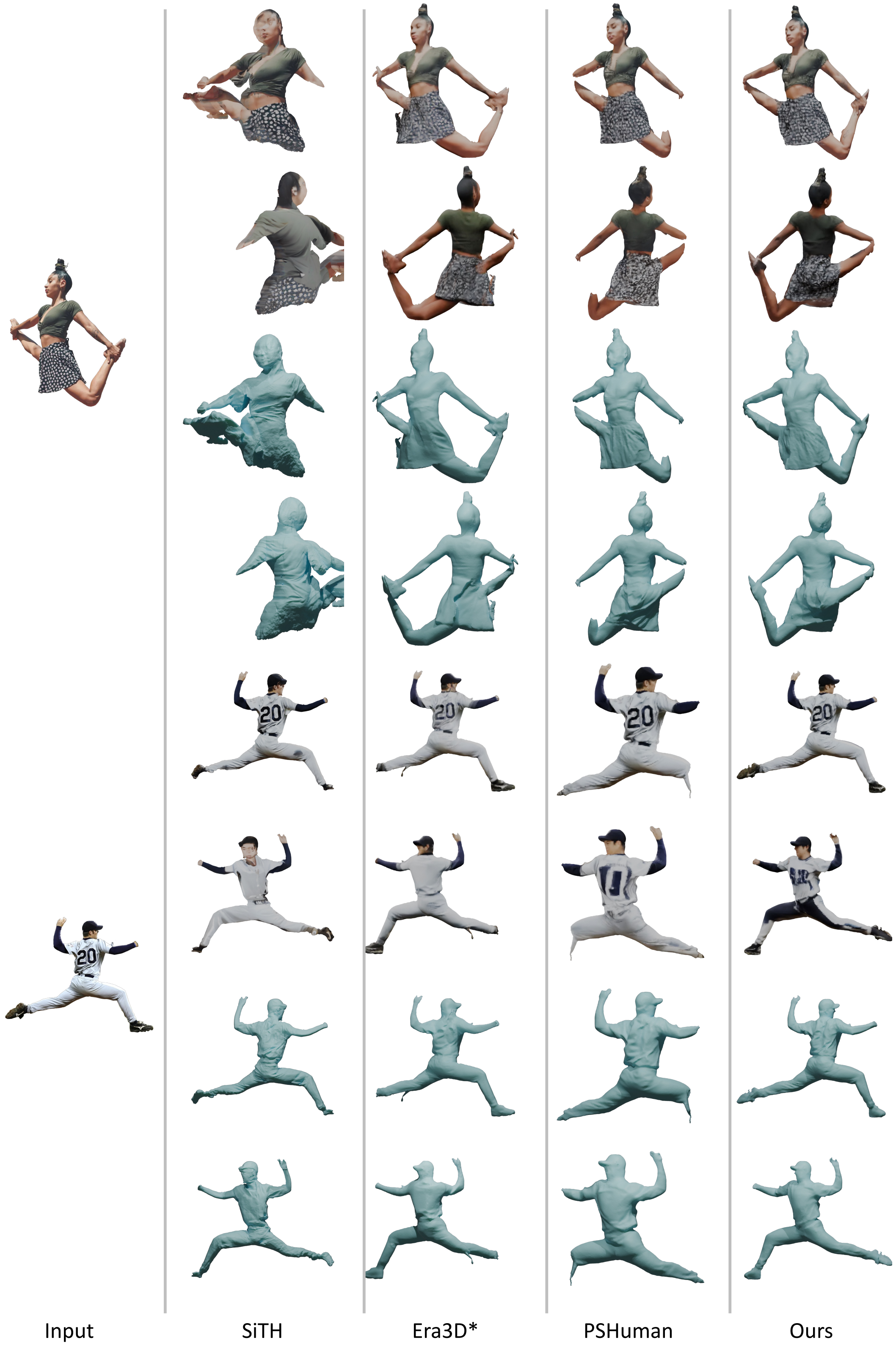

Figure 14: Additinoal qualitative evaluation on the internet-source images. Era3D* denotes Era3D fine-tuned on CustomHumans and THuman2.1 datasets.

Table 5: MIXAMORP dataset specification

Character	Animation	Description	Frame number(s)
Carla	Drop Kick	-	35, 46, 62
Carla	Start Plank	-	137
Claudia	Freehang Climb	-	47, 67
Claudia	Flying Knee Punch Combo	-	29, 79
Eric	Swing To Land	Swing Backflip To Crouched Land	26, 58
Eric	Standing Up	Sitting To Standing	41, 88
Henry	Situp To Idle	-	15, 49, 70
Henry	Female Standing Pose	On Left Leg, Right Hand...	1
Johanna	Twist Dance	-	163
Johanna	Jump Push Up	-	25
Johanna	Sitting Laughing	-	67
Johanna	Praying	...Prayer To Standing Up	1
Kumar	Rifle Turn And Kick	-	40, 48
Kumar	Dancing Twerk	-	179
Kumar	Crouch Turn Left 90	Turning 90 Degrees Left	6
Michael	Pain Gesture	-	20
Michel	Breakdance 1990	...Handstand Spin Start	1, 82, 100
Mira	Change Direction	-	25
Mira	Mma Kick	Mma Medium Kick	15, 22
Mira	Beckoning	-	26
Otto	Throw Grenade	...While In Prone Position	65
Otto	Run Backwards	...Backwards To Crouched Stop	37
Otto	Hurricane Kick	-	16
Otto	Grabbing Ammo	-	74
Sebastian	Pistol Kneeling Idle	-	1
Sebastian	Crawling	-	34
Sebastian	Dig And Plant Seeds	-	15, 70
Sheila	Shuffling	-	33
Sheila	Great Sword Slash	Great Sword Combo Slash	47, 55, 62
Sydney	Sword And Shield Attack	Sword And Shield High Attack	17, 26
Sydney	Running Jump	Jumping From A Sprint	7, 22
Tiffany	Samba Dancing	Afoxe Samba Reggae Dance	139
Tiffany	Stable Sword Inward Slash	-	5, 27
Toshiro	Martelo 2	-	17
Toshiro	Jump Attack	-	11, 27, 53
Victoria	Great Sword Crouching	...Sword Crouch To Block	10
Victoria	Chapa-Giratoria	-	61
Victoria	Jab Cross	Boxing Jab Cross Medium	22
Victoria	Jump	Jump In Place	35

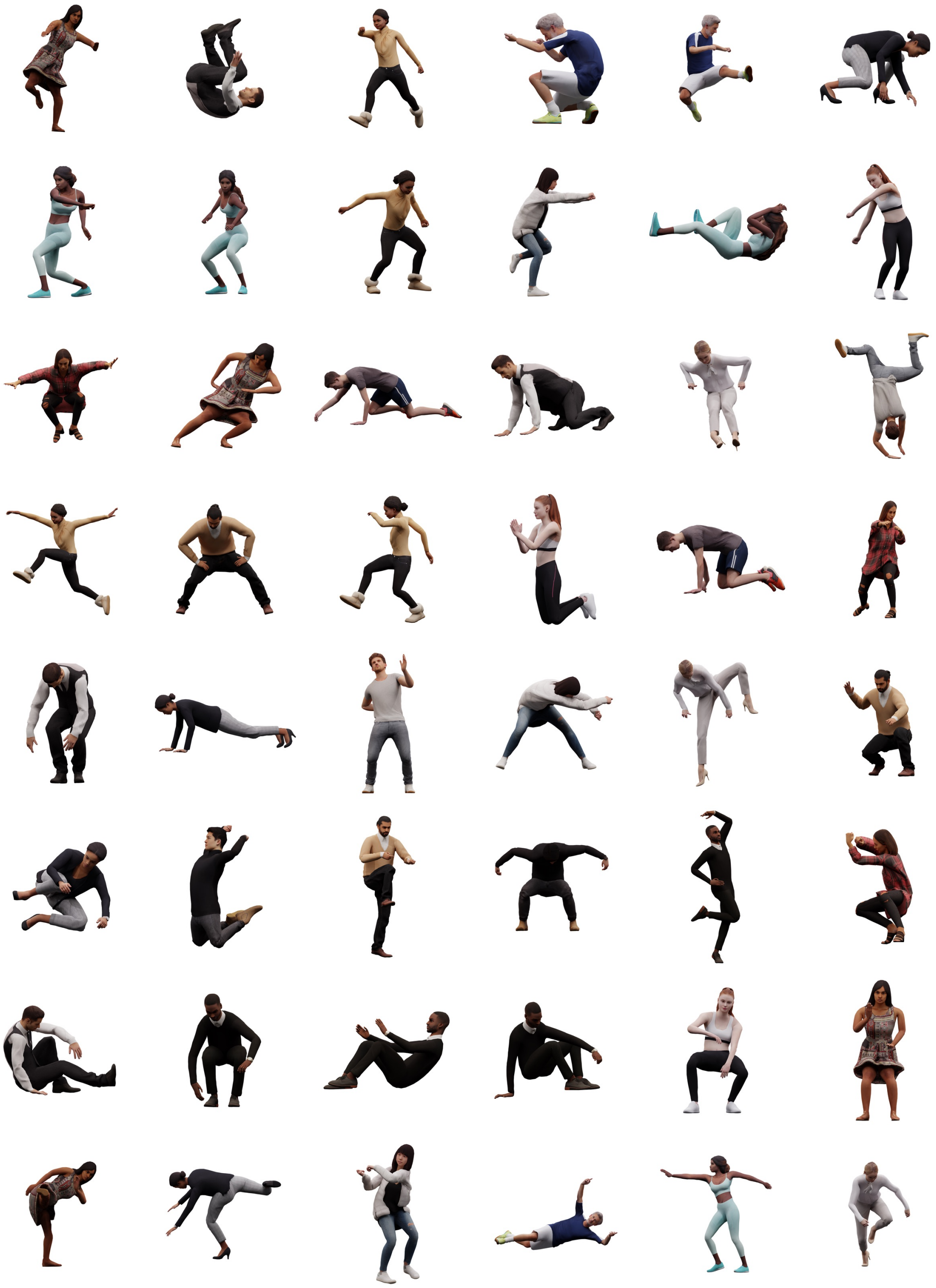

Figure 15: Representative visualizations for the MIXAMORP benchmark. The 48 meshes shown were randomly sampled from the complete dataset containing 60 meshes