BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion Supplementary Material

Michael J. Black[1,*] Priyanka Patel[1,*] Joachim Tesch[1,*] Jinlong Yang[2,*,†]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]Google

This document supplements the main text with (1) More details about the creation of the dataset. (2) More statistics about the dataset's contents. (3) More example images from the dataset. (4) Experimental results referred to in the main text. (5) Visual presentation of the qualitative results.

In addition to this document, please see the **Supplemental Video**, where the motions in the dataset are presented. The video, data, and related materials can be found at https://bedlam.is.tue.mpg.de/

BEDLAM: Definition

> noun
> *A scene of uproar and confusion: there was bedlam in the courtroom.*

The name of the dataset refers to the fact that the synthetic humans in the dataset are animated independently of each other and the scene. The resulting motions have a chaotic feel; please see the video for examples.

1. Dataset creation

Body shape diversity. The AGORA [24] dataset has 111 adult bodies in SMPL-X format [25]. These bodies mostly correspond to models with low BMI. Why do we use the bodies from AGORA? To create synthetic clothing we focused on creating synthetic versions of the clothed scans in AGORA. That is, we create "digital twins" of the AGORA scans. Our hope is that having 3D scans paired with simulated digital clothing will be useful for research on 3D clothing. Thus our 3D clothing is designed around AGORA bodies. Note that we do not make use of this property in BEDLAM but did this to enable future use cases. To increase diversity beyond AGORA, we sample an additional 80 male and 80 female bodies with BMI $>$ 30 from the CAESAR dataset [30].

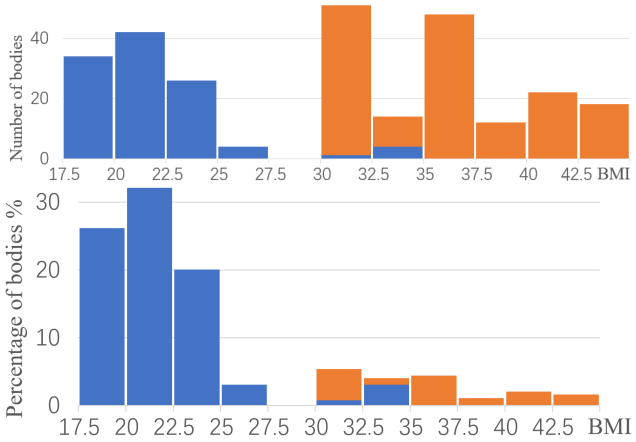

Figure 1. Body diversity in BEDLAM. Top: BMI distribution of the 271 different body shapes uses in BEDLAM. Bottom: BMI distribution in all rendered videos; 55009 in total. Blue bars represent bodies from the AGORA dataset, while orange bars represents high-BMI bodies from CAESAR dataset. BEDLAM uses both to cover a wide range of BMIs.

Note that the AGORA and CAESAR bodies are represented in gendered shape spaces using 10 shape components. When we render the images, we use these gendered bodies. For BEDLAM we use a gender-neutral shape space, enabling networks to automatically learn the appropriate body shape within this space, effectively learning to recognize gender. To make the ground truth shapes for BEDLAM in this gender-neutral space, we fit the gender-neutral model with 11 SMPL-X shape components to the gendered bodies. This is trivial since the meshes are in full correspondence. We use 11 shape components because, in the gender neutral space, the first component roughly captures the differences between male and female body shapes. Thus, adding one extra component means that the SMPL-X ground truth (GT) approximates the original gendered body shapes. There is some loss of fidelity but it is minimal; the V2V error between the rendered bodies and the GT bodies in neutral pose

*The authors contributed equally and are listed alphabetically.
†This work was performed when JL was at MPI-IS.

is 2.4mm.

Ideally, we want a diversity of body shapes, from slim to obese. Figure 1 shows the distribution of body BMIs in the training set. Specifically, we show the distribution of AGORA and CAESAR bodies, from which we sample. We also show the final distribution of BMIs in the training images.

Notice that the AGORA bodies are almost all slim. We add the CAESAR bodies to increase diversity and enable the network to predict high-BMI shapes. There is a dip in the distribution between 25-30 BMI. This happens to be precisely where the peak of the real population lies. Despite this lack of average BMIs, BEDLAM does a good job of predicting body shape, suggesting that it has learned to generalize.

Note that is it not clear what the right distribution for training is – one could mimic the distribution of a specific population or uniformly sample across BMIs. We plan to evaluate this and increase the diversity of the dataset; please check the project page for updates. Future work should also expand the types of bodies used to include children and people with diverse body types (athletes, little people, scoliosis, amputees, etc.).

Note that draping high-BMI models in clothing is challenging because the mesh self-intersects, causing failures of the cloth simulation. Future work could address this by automatically removing such intersections. Additionally, there is little motion capture data of obese people. So we need to retarget AMASS motions [17] to high-BMI subjects. But this is also problematic. Naive retargeting of motion from low-BMI bodies to high-BMI bodies results in interpenetration.

Here we use a simple solution to this problem. Given a motion sequence from AMASS, we first replace the original body shape with a high-BMI body. Then, we optimize the pose for each frame to minimize the body-body intersection using the code provided by TUCH [21]. Although this resolves interpenetration between body parts, it can create jittery motion sequences. As a remedy, we then smooth the jittery motion with a Gaussian kernel. Although this simple solution does not guarantee a natural motion without body-body interpenetration, it is sufficient to create a good amount of valid motion sequences for larger bodies. Future work should address the capture or retargeting of motion for high-BMI body shapes.

Skin tone diversity. Our skin tones were provided by Meshcapade GmbH and are categorized into several ethnic backgrounds, with skin-tone variety within each category. To generate BEDLAM subjects, we sample uniformly from the Meshcapade skins. This means the final renders are sampled with the following representations

- African 20%,
- Asian 24%,
- Hispanic 6%,
- Indian 20%,
- Mideast 6%,
- South East Asian 10%,
- White 14%.

The same proportions hold in the training, validation and test sets.

Figure 2. Clothing deformation is well modeled by physics-based simulation.

Motion sampling. Due to the imbalanced distribution of motions in AMASS, we use the motion labels from BABEL [27] to sample the motions for a wide and even coverage of the motion space. After visualizing the motions in each labelled category, we manually assign the number of motions sampled from each category. Specifically, we sample 64 sequences for motions such as “turn”, “cartwheel”, “bend”, “sit ”, “touch ground”, etc. We sample 4 sequences from motion labels containing less pose variation, such as “draw”, “smell”, “lick”, “listen ”, “look”, etc. We do not sample any sequences from labels indicating static poses, for example, “stand”, “a pose”, and “t pose”. For the remaining motion labels, we sample 16 random sequences from each. Each sampled motion sequence lasts from 4 to 8 seconds.

Clothing. Our outfits are designed to reflect real-world clothing complexity. We have layered garments and detailed structures such as pleats and pockets. We also have open jackets and many wide skirts, which usually have large deformation under different body motion. These deformations can only be well modeled with a physics-based simulation. See Fig. 2 for examples.

Putting multiple people in the scene. For each sequence we randomly select between 1 and 10 subjects. For each subject a random animation sequence is selected. The shortest animation sequence determines the image sequence

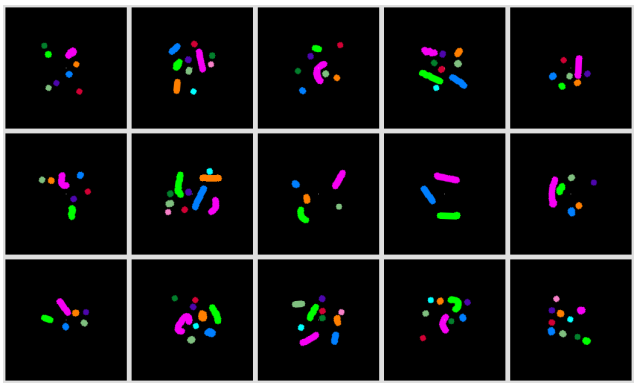

Figure 3. Examples of animation ground trajectories. Top-view pelvis trajectories, color coded by subject. These trajectories are automatically placed so that the bodies do not collide. Here, 15 sample sequences are shown with varying numbers of subjects.

length to ensure that there are no "frozen" body poses. We then pick a random sub-motion of the desired sequence length from each body motion in the sequence. Next the body motions are placed in a desired target area of the scene at a randomized position with a randomized camera yaw. To avoid overlapping body motions and collisions with the 3D environment, we use 2D binary ground plane occupancy masks of the pelvis location for each randomly placed motion. The order of motion placement is determined by the ground plane pelvis coverage bounding box. This ensures that walking motions, which are challenging to place in a limited space, have the maximum free ground space available before more constrained motions fill the remaining space; cf. [1]. Generated root trajectories can be seen in Fig. 3. This is a simple strategy (cf. [1]) and future work should explore the generation or placement of motions that make more sense together and with respect to the scene. One direction would use MIME [38] to take human motions and produce 3D scenes that are consistent with them.

Additional limitations: Hair and shadows. Designing high-quality hair assets requires experienced artists. Here we used a commercial hair solution based on "hair cards"; these are simpler than strand-based methods. The downside is that they require the use of temporal accumulation buffers in the deferred rendering system. This can introduce ghosting artefacts when rendering fast motions at low frame rates. We also observed hair shader illumination issues under certain conditions. When used with the new real-time global illumination system (Lumen) in Unreal Engine 5 (UE5), some hairstyles exhibit a strong hue shift. Also, the number of hair colors that we have is limited. When used in the HDRI environments, with ray traced HDRI shadows enabled, most hairstyles turn black. For this reason we do not use ray traced HDRI shadows in the HDRI environment renders, though the 3D scenes do have cast shadows. Adding ground contact shadows to the HDRI scenes would require the use of a separate ground shadow caster render pass to composite the shadow into the image. We have not pursued this because we plan to upgrade the hair assets to remove these issues for future releases of the dataset.

Other body models. BEDLAM is designed around SMPL-X but many methods in the field use SMPL [15]. In particular, most, if not all, current methods that process video sequences are based on SMPL and not SMPL-X. We will provide the ground truth in SMPL format as well for backward compatibility. We also plan to support other body models like GHUM [36] or SUPR [22] in the future.

Additional ground truth data: Depth maps and semantic segmenation. Since BEDLAM is rendered with UE5, we can render out more than RGB images. In particular, we render depth maps and segmentation masks as illustrated in Fig. 4. The segmentation information includes semantic labels for hair, clothing and skin. With these additional forms of ground truth, BEDLAM can be used to train and evaluate methods that regress depth from images, fit bodies to RGB-D data, perform semantic segmentation, etc.

Assets. We will make available the rendered images and the SMPL-X ground truth. We also release the 3D clothing and clothing textures as well as the skin textures. We also will make available the process to create more data. All assets used are described in Table 1. The table provides a "shopping list" to recreate BEDLAM. The only asset that presents a problem for recreating BEDLAM is the hair since new licenses of the the hair assets prohibit training of neural networks (we acquired the data under an older license). This motivates us to develop new hair assets with an unrestricted license. More information about how to create new data is provided on the project website.

2. Comparison to other datasets

Table 2 compares synthetic datasets mentioned in the related work section of the main paper. Here we only survey methods that provide images with 3D ground truth; this excludes datasets focused solely on 3D clothing modeling. Some of the listed datasets are not public but we include them anyway and some information is not provided in the publications ("unk." in the table).

Methods vary in terms of the number of subjects, from a handful of bodies to over 1000 in the case of Ultrapose. Ultrapose, however, is not guaranteed to have realistic bodies and the dataset is biased towards mostly thin Asian bodies. The released dataset also has blurred faces. The number of frames also varies significantly among datasets. To get a

Figure 4. Additional ground truth: Depth maps and semantic segmentation masks. The segmentation maps are color coded for each individual and each material type (hair, clothing, skin).

Asset Type	Name	Source
Body Texture	Various	Meshcapade GmbH, https://meshcapade.com
Hair	Prime Hairstyles	Reallusion, https://www.reallusion.com/ContentStore/Character-Creator/Pack/Prime-hairstyles/
Hair	Trendy Hairstyles for Men Vol. 1	Reallusion, https://www.reallusion.com/ContentStore/Pack/universal-hairstyles-vol-1
Hair	Trendy Hairstyles for Men Vol. 2	Reallusion, https://www.reallusion.com/ContentStore/Pack/universal-hairstyles-vol-2
Environment - HDRI	Various free HDRIs	Poly Haven, CC0 1.0 Universal Public Domain Dedication, https://polyhaven.com/hdris
Environment - 3D	ArchViz User Interface 3	https://www.unrealengine.com/marketplace/en-US/product/archviz-user-interface-3
Environment - 3D	Big Office	https://www.unrealengine.com/marketplace/en-US/product/big-office
Environment - 3D	High School Basketball Gym	https://www.unrealengine.com/marketplace/en-US/product/high-school-basketball-gym-day-night-afternoon-midnight-lighting
Environment - 3D	Sports Stadium	https://www.unrealengine.com/marketplace/en-US/product/sports-stadium
Environment - 3D	Suburb Neighborhood House Pack	https://www.unrealengine.com/marketplace/en-US/product/suburb-neighborhood-house-pack-modular

Table 1. Third-party assets used for rendering BEDLAM. All 3D environments are from the Unreal Marketplace.

sense of the diversity of images, one must multiply the number of frames by the average number of subjects per image (Sub/image).

The methods vary in how images are generated. The majority composite a rendered 3D body onto an image background. This has limited realism. Human3.6M has mixed reality data in which simple graphics characters are inserted into real scenes using structure from motion. Mixed/composite methods capture images of real people with a green screen in a multi-camera setup. They can then get pseudo-ground tuth and composite the original images on new backgrounds. In the table, "rendered" means that the synthetic body is rendered in a scene (HDRI panorama or 3D model) with reasonable lighting. These are the most realistic methods.

Clothing in previous datasets takes several forms. The simplest is a texture map on the SMPL body surface (like in SURREAL [34]). Some methods capture real clothing or use scans of real clothing. Another class of methods uses commercial "rigged" models with rigged clothing. This type of clothing lacks the realism of physics simulation. Most methods that do physics simulation use a very limited number of garments (often as few as 2) due to the complexity and cost.

It is hard to get good, comparable, data about motion diversity in these datasets. Here we list numbers of motions gleaned from the papers but these are quite approximate. Some of the low numbers describe classes of motions that may be repeated with some unknown number of variations. At the same time, some of the larger numbers may lack diversity. With BEDLAM, we are careful to sample a diverse set of motions.

For comparison with real-image datasets, 3DPW contains 60 sequences captured with a moving camera, with roughly 51K frames, and 7 subjects in a total of 18 clothing styles. With roughly 2 subjects per frame, this gives around 100K unique bounding boxes. Human3.6M training data has 1,464,216 frames captured by 4 static cameras at 50 fps, which means there are 366K unique articulated poses. If one reduces the frame rate to 30 fps, that gives roughly 220K bounding boxes of 5 subjects performing 15 different types of motions. We observe that the total number of frames is less important than the diversity of those frames in terms of scene, body, pose, lighting, and clothing.

3. Implementation Details

BEDLAM-CLIFF-X. Since most HPS methods output SMPL bodies, we focus on that in the main paper and describe the SMPL-X methods here. Specifically, we use BEDLAM hand poses to train a full body network called BEDLAM-CLIFF-X. For this, we train a separate hand network on hand crops from BEDLAM with an HMR architecture but replace SMPL with the MANO hand [31], which is compatible with SMPL-X. We merge the body pose output $\theta_b \in \mathbb{R}^{22\times3}$ from BEDLAM-CLIFF (see Sec. 4.1 of the main paper) and hand pose output $\theta_h \in \mathbb{R}^{16\times3}$ from the hand network to get the full body pose with articulated hands $\theta_{fb} \in \mathbb{R}^{55\times3}$. The face parameters, θ_{jaw}, θ_{leye} and θ_{reye} are kept as neutral. Since both BEDLAM-CLIFF and the hand network output different wrist poses, we cannot

Dataset	#Sub	#Frames	Image	Subj/image	Clothing	Motion	Ground truth
3D HUMANS-Train [7]	19	50K	composite	1	captured	>15	SMPL
SURREAL [34]	145	≈6.5M	composite	1	texture	> 2000	SMPL
Human3.6M [8]	few	7.5K	mixed reality	1	rigged	unk.	3D joints
MPI-INF-3DHP-Train [18]	8	>1.3M	mixed/composite	1	real	8+	3D joints
MuCo-3DHP [19]	8	≈400K	mixed/composite	1-4	real	8	3D joints
Daněček et al. [5]	10	unk.	rendered (simple)	1	physics	20 min	unk.
Liang and Lin [12]	100	128K	composite	1	physics	5 seqs	SMPL
BCNet (a) [9]	285	13K	composite	1	rigged	unk.	SMPL
BCNet (b) [9]	3048	17K	composite	1	static physics	55	SMPL
Liu et al. [14]	unk.	3M	composite	1	physics	5k	SMPL
Ultrapose [37]	>1000	≈500K	composite	1	physics	n/a	dense points
3DPeople [26]	80	≈2.5M	composite	1	rigged	70	3D joints
HSPACE [1]	100	1M	rendered	5 avg.	rigged (100)	100	GHUM
GTA-Human [3]	>600	≈ 1.4M	game	1	rigged	20K	SMPL
AGORA [24]	>350	≈18K	rendered	5-15	scans	n/a	SMPL-X, SMPL
BEDLAM (ours)	217	380K	rendered	1-10	physics (110)	2311	SMPL-X

Table 2. Comparison of synthetic human datasets that provide images with 3D human pose annotations. See text.

merge them directly. Hence, we train a small regressor R_{fb} to combine them.

Specifically, we define the body pose $\theta_b = \{\hat{\theta}_b, \theta_{elbow}, \theta^b_{wrist}\}$ and and hand pose $\theta_h = \{\theta^h_{wrist}\ \theta_{fingers}\}$, where $\hat{\theta}_b \in \mathbb{R}^{20\times3}$ represents the first 20 pose parameters of SMPL-X. R_{fb} takes global average pooled features as well as θ_b and θ_h from the BEDLAM-CLIFF and hand networks, and outputs $\theta_{fb} = \{\hat{\theta}_b, \theta_{elbow} + \Delta_{elbow}, \theta^b_{wrist}+\Delta_{wrist}, \theta_{fingers}\}$. Basically, R_{fb} learns an update of the elbow and wrist pose from the body network using information from both the body and hand network. Since we learn only an update on the wrist pose generated by the body network, this prevents the unnatural bending of the wrists. Similar to BEDLAM-CLIFF, to train BEDLAM-CLIFF-X, we use a combination of MSE loss on model parameters, projected keypoints, 3D joints, and an L1 loss on 3D vertices. All other details can be found the code (see project page).

Data augmentation. A lot of data augmentation is included during training, including random crops, scale, different kinds of blur and image compression, brightness and contrast modification, noise addition, gamma, hue and saturation modification, conversion to grayscale, and downscaling using [2].

4. Supplemental experiments

4.1. Ablation of training data and backbones

Table 3 expands on Table 3 from the main paper, providing the full set of dataset ablation experiments. The key takeaways are: (1) training with a backbone pretrained on the 2D pose-estimation task on COCO produces the best results, (2) training from scratch on BEDLAM does not work as well as either pre-training on ImageNet or COCO, (3) training only on BEDLAM is better than training only on AGORA, (4) training on BEDLAM+AGORA is consistently better than using either alone (note that both are synthetic), (5) one can get by with using a fraction of BEDLAM (50% or even 25% gives good performance), but training error continues to decrease up to 100%. All of this suggest that there is still room for improvement in the synthetic data in terms of variety.

Method	Dataset	Backbone	Crops %	PA-MPJPE	MPJPE	PVE
HMR	B+A	scratch	100	67.9	108.8	129.0
HMR	B+A	ImageNet	100	57.3	91.7	108.8
HMR	B+A	COCO	100	47.6	79.0	93.1
CLIFF	B+A	scratch	100	61.7	96.5	115.0
CLIFF	B+A	ImageNet	100	51.8	82.1	96.9
CLIFF	B+A	COCO	100	47.4	73.0	86.6
HMR	B	COCO	5	55.8	86.9	104.3
HMR	B	COCO	10	55.5	85.7	102.9
HMR	B	COCO	25	53.9	83.9	100.4
HMR	B	COCO	50	53.8	81.1	97.3
HMR	B+A	COCO	100	47.6	79.0	93.1
CLIFF	B	COCO	5	54.0	80.8	96.8
CLIFF	B	COCO	10	53.8	79.9	95.7
CLIFF	B	COCO	25	52.2	77.7	93.6
CLIFF	B	COCO	50	51.0	76.3	91.1
CLIFF	B+A	COCO	100	47.4	73.0	86.6
HMR	A	COCO	100	58.3	94.9	109.0
HMR	B	COCO	100	51.2	80.6	96.1
HMR	B+A	COCO	100	47.6	79.0	93.1
CLIFF	A	COCO	100	54.0	88.0	101.8
CLIFF	B	COCO	100	50.5	76.1	90.6
CLIFF	B+A	COCO	100	47.4	73.0	86.6

Table 3. Ablation experiments on 3DPW. B denotes BEDLAM and A denotes AGORA. Crops % only applies to BEDLAM.

4.2. Ablation on losses

To understand which loss terms are important, we perform an ablation study on standard losses used in training

HPS methods including L_{SMPL}, L_{j3d}, L_{j2d}, L_{v3d}, L_{v2d}. Individual losses are described here and the ablation on them is reported in Table 4.

$$L_{\text{SMPL}} = \|\hat{\theta} - \theta\| + \|\hat{\beta} - \beta\|$$

$$L_{j3d} = \|\hat{\mathcal{J}} - \mathcal{J}\|$$

$$L_{j2d} = \|\hat{j} - j\|$$

$$L_{v3d} = \|\hat{\mathcal{V}} - \mathcal{V}\|$$

$$L_{v2d} = \|\hat{v} - v\|$$

$\hat{x}$ denotes the ground truth for the corresponding variable x and $\|\cdot\|$ is the type of loss that can be L1 or L2. For shape we always use L1 norm. $\mathcal{J}$, $\mathcal{V}$, β and θ denote the 3D joints, 3D vertices, shape and pose parameters of SMPL-X model respectively. j and v denote the 2D joints and vertices projected into the full image using the predicted camera parameters similar to [11]. θ is predicted in a 6D rotation representation form [40] and converted to a 3D axis-angle representation when passed to SMPL-X model. Since we set the hand poses to neutral in BEDLAM-CLIFF, we use only the first 22 pose parameters in the training loss. We use a subset of BEDLAM training data for this ablation study. Note that, to compute L_{v2d} we use a downsampled mesh with 437 vertices, computed using the downsampling method in [28]. We find this optimal for training speed and performance. Since the downsampling module samples more vertices in regions with high curvature, it helps preserve the body shape and we can store the sampled vertices directly in memory without the need to load them during training. We include a 2D joints loss in all cases as it is necessary to obtain proper alignment with the image.

As shown in Table 4, L_{j3d} or L_{v3d} alone do not provide enough supervision for training. Similar to [23] we find that L_{SMPL} provides stronger supervision reducing the loss by a large margin when used in combination with L_{v3d} and L_{j3d}. Surprisingly, we find that including L_{v2d} makes the performance slightly worse. A plausible reason for this could be that using L_{v2d} provides high weight on aligning the predicted body to the image but the mismatch between the ground truth and estimated camera used for projection during inference makes the 3D pose worse, thus resulting in higher 3D error. We suspect that L_{v2d} could provide strong supervision in the presence of a better camera estimation model; this is future work.

We also experiment with two different types of losses, L1 and MSE and find that L1 loss yields lower error on the 3DPW dataset as shown in Table 4. However, Table 5 shows that the model using L1 loss performs worse when estimating body shape on the SSP and HBW datasets compared to the model using MSE loss. This discrepancy may be attributed to the L1 loss treating extreme body shapes as outliers, thereby learning only average body shapes. Since the 3DPW dataset does not have extreme body shapes, it benefits from the L1 loss. Consequently, we opted to use the MSE loss for our final model and all results reported in the main paper. Note that L_{j3d} or L_{v3d} alone is worse with L1 loss compared to MSE loss.

Losses	Type	PAMPJPE	MPJPE	MVE
L_{j3d}	MSE	59.1	86.1	105.1
L_{v3d}	MSE	56.2	83.4	96.7
L_{SMPL}	MSE	51.3	83.8	96.7
L_{SMPL} + L_{j3d}	MSE	48.5	76.0	89.6
L_{SMPL} + L_{v3d}	MSE	48.2	74.7	87.9
L_{SMPL} + L_{v3d} + L_{j3d}	MSE	**47.6**	**74.2**	**87.2**
L_{SMPL} + L_{v3d} + L_{j3d} + L_{v2d}	MSE	48.7	74.4	87.6
L_{j3d}	L1	59.4	85.7	114.6
L_{v3d}	L1	72.5	97.4	111.6
L_{SMPL}	L1	50.6	83.6	96.0
L_{SMPL} + L_{j3d}	L1	46.9	74.7	87.6
L_{SMPL} + L_{v3d}	L1	48.8	76.2	88.8
L_{SMPL} + L_{v3d} + L_{j3d}	L1	**46.9**	**73.0**	**86.0**
L_{SMPL} + L_{v3d} + L_{j3d} + L_{v2d}	L1	47.4	73.5	86.8

Table 4. **Ablation of different losses.** Error on 3DPW in mm.

Loss type	SSP-3D	HBW				
	PVE-T-SC	Height	Chest	Waist	Hips	$P2P_{20k}$
L1	15.1	51	73	97	64	22
MSE	14.2	51	69	88	62	22

Table 5. **Losses.** The use of L2 or L1 losses are explored for shape estimation accuracy using BEDLAM-CLIFF: error on HBW [21] and SSP-3D [33] in mm.

4.3. Ablation of dataset attributes

We also perform an ablation study by varying different dataset attributes. We generated 3 different sets of around 180K images by varying the use of different assets. Keeping the scenes and the motion sequences exactly the same, we experiment by ablating hair and then further replacing the cloth simulation with simple cloth textures. We use a backbone pretrained with either COCO [13] or ImageNet and study the performance on 3DPW [35]. When using the ImageNet backbone, we find that training with clothing simulation leads to better accuracy than training with clothing texture mapped onto the body. Adding hair gives a modest improvement in MPJPE and MVE. Surprisingly, with the COCO backbone, the difference in the training data makes less difference. Still, clothing simulation is consistently better than just using clothing textures. It is likely that the backbone pretrained on a 2D pose estimation task using COCO is already robust to clothing and hair. As mentioned above, however, our hair models are not ideal and not as diverse as we would like. Future work, should explore whether more diverse and complex hair has an impact.

Dataset attribute	Backbone	PAMPJPE	MPJPE	MVE
Simulation + Hair	ImageNet	65.6	101.8	120.8
Simulation	ImageNet	66.3	104.5	124.5
Texture	ImageNet	72.2	116.1	136.7
Simulation + Hair	COCO	51.6	77.8	92.4
Simulation	COCO	51.6	78.7	93.0
Texture	COCO	54.3	80.8	96.0

Table 6. **Ablation of different dataset attributes.** Error on 3DPW in mm. See text.

Method	H3.6M		3DPW		
	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PVE
CLIFF [11]	32.7	47.1	-	-	-
CLIFF†*	39.4	62.9	43.6	68.8	82.1
CLIFF†* w/o H3.6M	56.1	89.6	44.4	68.9	82.3
BEDLAM-HMR	51.7	81.6	47.6	79.0	93.1
BEDLAM-CLIFF	50.9	70.9	46.6	72.0	85.0

Table 7. **Impact of training without Human3.6M on Human3.6M and 3DPW.** CLIFF†* is the same model as Table 1 in main paper.

4.4. Experiment on Human3.6M

We also evaluate our method on the Human3.6M dataset [8] by calculating MPJPE and PA-MPJPE on 17 joints obtained using the Human3.6M regressor on vertices. Previous methods have used Human3.6M training images when evaluating on the test set. Specifically, CLIFF [11] and our re-implementation, CLIFF†*, both use Human3.6M data for training and, consequently get low errors on Human3.6M test data. Note that our implementation does not get as low an error as reported in [11] despite the fact that we match their performance on 3DPW and RICH (see main paper).

To ensure a fair comparison and to measure the generalization of the methods, we trained a version of CLIFF (CLIFF†* w/o H3.6M) using 3D datasets MPI-INF-3DHP, 3DPW and 2D datasets COCO and MPII but excluding Human3.6M, following the same settings as BEDLAM-CLIFF. The results in Tab. 7 demonstrate that BEDLAM-CLIFF outperforms CLIFF when Human3.6M is not included in training. This is another confirmation of the results in the main paper showing that BEDLAM-CLIFF has better generalization ability than CLIFF. Without using Human3.6M in training, BEDLAM-HMR is also better than CLIFF on Human3.6M.

Note that this experiment illustrates how training on Human3.6M is crucial to getting low errors on that dataset. The training and test sets are similar (same backgrounds and similar conditions) meaning that methods trained on the dataset can effectively over-fit to it. This can be seen by comparing CLIFF†* with CLIFF†* w/o H3.6M. Training on Human3.6M significantly reduces error on Human3.6M without reducing error on 3DPW.

4.5. SMPL-X experiments on the AGORA dataset

AGORA is interesting because it is one of the few datasets with SMPL-X ground truth. Table 8 evaluates methods that estimate SMPL-X bodies on the AGORA dataset. The results are taken from the AGORA leaderboard. BEDLAM-CLIFF-X does particularly well on the face and hands. Since the BEDLAM training set contains body shapes sampled from AGORA, it gives BEDLAM-CLIFF-X an advantage over methods that are not fine-tuned on the AGORA training set (bottom section of Tab. 8). Consequently, we also compare a version of BEDLAM-CLIFF-X that is trained only on the BEDLAM training set. This still outperforms all the methods that were not trained using AGORA (top section of Tab. 8). Please see Figure 6 for qualitative results.

4.6. SMPL-X experiments on BEDLAM

For completeness, Tab. 9 shows that BEDLAM-CLIFF-X outperforms recent SOTA methods that estimate SMPL-X on the BEDLAM test set. Not surpisingly, our method is more accurate by a large margin. Note, however, that the prior methods are not trained on the BEDLAM training data. We follow a similar evaluation protocol as [24]. Since the hands are occluded in a large number of frames, we use MediaPipe [16] to detect the hands and evaluate hand accuracy only if they are visible. To detect individuals within an image during evaluation, we use the detector that is included in the respective method's demo code. In cases where the detector is not provided, we use [29], the same detector use by BEDLAM-CLIFF-X. Please see Fig. 6 for qualitative results.

5. Qualitative Comparison

Figure 5 provides a qualitative comparison between PARE [10], CLIFF [11] (includes 3DPW training) and BEDLAM-CLIFF (only synthetic data). We show results on both RICH (left two) and 3DPW (right two). We render predicted bodies overlaid on the image and in a side view. In the side view, the pelvis of the predicted body is aligned (translation only) with the ground truth body. Note that, when projected into the image, all methods look reasonable and relatively well aligned with the image features. The side view, however, reveals that BEDLAM-CLIFF (bottom row) predicts a better aligned body pose with the ground truth body in 3D despite variation in the cameras, camera angle, and frame occlusion. Also, please notice that BEDLAM-CLIFF produces more natural leg poses in the case of occlusion compared to the other methods as shown in columns 1, 3 and 4 of Fig. 5

We also provide qualitative results of BEDLAM-CLIFF-X on 3DPW and the RICH dataset in Fig. 7. In this case,

Method	MVE				MPJPE			
	FB	B	F	LH/RH	FB	B	F	LH/RH
SMPLify-X [25]	236.5	187.0	48.9	48.3/51.4	231.8	182.1	52.9	46.5/49.6
ExPose [4]	217.3	151.5	51.1	74.9/71.3	215.9	150.4	55.2	72.5/68.8
Frankmocap [32]		168.3		54.7/55.7		165.2		52.3/53.1
PIXIE [6]	191.8	142.2	50.2	49.5/49.0	189.3	140.3	54.5	46.4/46.0
BEDLAM-CLIFF-X	**131.0**	**96.5**	**25.8**	**38.8/39.0**	**129.6**	**95.9**	**27.8**	**36.6/36.7**
Hand4Whole+ [20]	135.5	90.2	41.6	46.3/48.1	132.6	87.1	46.1	44.3/46.2
PyMAF+ [39]	125.7	84.0	35.0	44.6/45.6	124.6	83.2	37.9	42.5/43.7
BEDLAM-CLIFF-X+	**103.8**	**74.5**	**23.1**	**31.7/33.2**	**102.9**	**74.3**	**24.7**	**29.9/31.3**

Table 8. **SMPL-X methods on the AGORA test set.** + denotes methods include AGROA training set. FB is full-body, B is body only, F is face, and LH/RH are the left and right hands respectively.

Method	NMVE		NMJE		MVE				MPJPE			
	FB	B	FB	B	FB	B	F	LH/RH	FB	B	F	LH/RH
PyMAF-X [39]	172.1	123.6	167.2	120.1	161.8	117.4	50.3	40.5/42.6	157.2	114.1	51.6	38.2/39.7
Hand4Whole [20]	178.8	119.1	176.2	117.6	168.1	112.0	59.7	52.8/55.8	165.7	110.5	63.7	50.0/52.0
PIXIE [6]	160.0	107.2	154.8	103.5	150.4	100.8	51.4	47.2/50.2	145.6	97.3	55.4	43.6/46.0
BEDLAM-CLIFF-X	101.7	65.6	99.0	64.7	95.6	61.7	29.9	35.7/36.2	93.1	60.8	30.5	33.2/33.3
BEDLAM-CLIFF-X+	**93.4**	**61.2**	**92.5**	**60.4**	**87.8**	**56.8**	**27.3**	**31.9/33.9**	**87.0**	**57.5**	**28.0**	**29.5/31.1**

Table 9. **SMPL-X methods on the BEDLAM test set.** Comparison of SOTA methods on the BEDLAM test set. + denotes methods include AGROA training set.

we also estimate the SMPL-X hand poses. All multi-person results are generated by running the method on individual crops found by a multi-person detector [29].

References

[1] Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. HSPACE: Synthetic parametric humans animated in complex environments. *arXiv*, 2112.12867, 2021. 3, 5

[2] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 5

[3] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, and Ziwei Liu. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021. 5

[4] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, volume 12355, pages 20–40, 2020. 8

[5] R. Daněček, E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. DeepGarment: 3D garment shape estimation from a single image. *Comput. Graph. Forum*, 36(2):269–280, may 2017. 5

[6] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pages 792–804, 2021. 8

[7] Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3D human shape estimation from single images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2232–2241, 2019. 5

[8] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2013. 5, 7

[9] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. BCNet: Learning body and cloth shape from a single image. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX*, pages 18–35, 2020. 5

[10] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 7

[11] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, 2022. 6, 7

[12] Junbang Liang and Ming C Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4352–4362, 2019. 5

Figure 5. Qualitative results on RICH (left two columns) and 3DPW (right two columns). RGB images (row 1), PARE front (row 2), PARE side (row 3), CLIFF front (row 4), CLIFF side (row 5), BEDLAM-CLIFF front (row 6), BEDLAM-CLIFF side (row 7). Ground truth body is in blue and predicted body is in pink. The BEDLAM-CLIFF predicted 3D body is better aligned with ground truth in both front and side views despite wide camera variation or frame occlusion.

Figure 6. BEDLAM-CLIFF-X results on the AGORA-test (top 4 rows) and the BEDLAM-test images (bottom 2 rows).

Figure 7. BEDLAM-CLIFF-X results on 3DPW-test (top 2 rows) and RICH-test (bottom 2 rows) images. Note the hand poses and that the body shapes are appropriately gendered.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755, 2014. 6

[14] Jian Liu, Naveed Akhtar, and Ajmal Mian. Temporally coherent full 3D mesh human pose recovery from monocular video. *arXiv preprint arXiv:1906.00161*, 2019. 5

[15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 3

[16] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. 7

[17] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 2

[18] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 5

[19] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *3DV*, 2018. 5

[20] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. 8

[21] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, 2021. 2, 6

[22] Ahmed A. A. Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. SUPR: A sparse unified part-based human representation. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, Oct. 2022. 3

[23] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 6

[24] Priyanka Patel, Chun-Hao Paul Huang, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021. 1, 5, 7

[25] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 8

[26] Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3DPeople: Modeling the Geometry of Dressed Humans. In *International Conference in Computer Vision (ICCV)*, 2019. 5

[27] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021. 2

[28] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–720, 2018. 6

[29] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 7, 8

[30] Kathleen M. Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoeferlin, and Dennis Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Technical Report AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory, 2002. 1

[31] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017. 4

[32] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 8

[33] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3D human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, 2020. 6

[34] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017. 4, 5

[35] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Eur. Conf. Comput. Vis.*, 2018. 6

[36] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 3

[37] Haonan Yan, Jiaqi Chen, Xujie Zhang, Shengkai Zhang, Nianhong Jiao, Xiaodan Liang, and Tianxiang Zheng. Ultrapose: Synthesizing dense pose with 1 billion points by human-body decoupling 3d model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10891–10900, 2021. 5

[38] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[39] 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. *arXiv preprint arXiv:2207.06400*, 2022. 8

[40] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 6