BEDLAM2.0: Synthetic Humans and Cameras in Motion

Joachim Tesch[1] **Giorgio Becherini**[1] **Prerana Achar**[1] **Anastasios Yiannakidis**[1]
Muhammed Kocabas[2] **Priyanka Patel**[2] **Michael J. Black**[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Meshcapade GmbH

Abstract

Inferring 3D human motion from video remains a challenging problem with many applications. While traditional methods estimate the human in image coordinates, many applications require human motion to be estimated in world coordinates. This is particularly challenging when there is both human and camera motion. Progress on this topic has been limited by the lack of rich video data with ground truth human and camera movement. We address this with BEDLAM2.0, a new dataset that goes beyond the popular BEDLAM dataset in important ways. In addition to introducing more diverse and realistic cameras and camera motions, BEDLAM2.0 increases diversity and realism of body shape, motions, clothing, hair, and 3D environments. Additionally, it adds shoes, which were missing in BEDLAM. BEDLAM has become a key resource for training 3D human pose and motion regressors today and we show that BEDLAM2.0 is significantly better, particularly for training methods that estimate humans in world coordinates. We compare state-of-the art methods trained on BEDLAM and BEDLAM2.0, and find that BEDLAM2.0 significantly improves accuracy over BEDLAM. For research purposes, we provide the rendered videos, ground truth body parameters, and camera motions. We also provide the 3D assets to which we have rights and links to those from third parties.

1 Introduction

The BEDLAM dataset [6] was the first synthetic dataset of 3D human motion of sufficient realism and complexity that synthetic data alone could be used to train a state-of-the-art (SOTA) method for estimating 3D human shape and pose (HPS) from images. Since its introduction, BEDLAM has become a standard dataset for supervised training of HPS regression methods. Despite this success, BEDLAM has several key limitations that hold back the field. Key among these is that BEDLAM uses limited camera focal lengths and camera motions. Here we address these limitations and provide a significantly richer dataset appropriate for end-to-end training of HPS methods that estimate humans in world coordinates. Beyond richer camera motions, the BEDLAM2.0 dataset addresses several other important limitations that improve its diversity and realism.

Beyond the camera motions, BEDLAM2.0 goes beyond B1 in the following ways:

- We significantly expand the range of body shapes with more high-BMI bodies.
- We provide more varied and realistic 3D hair that is adapted to individual head shapes.
- We add widely varied shoes, which are completely missing in BEDLAM. This includes defining the sole thickness, making foot-ground contact more realistic.
- We include more 3D clothing outfits and grade many outfits into standard sizes. In comparison to B1, this lets us dress diverse body shapes in realistic clothing.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

- B2 also includes more 3D scenes, more complex human motions, and longer motions, increasing the diversity of the dataset.

The most significant upgrade involves the cameras. Specifically, we define cameras that cover the range of focal lengths seen in real images and videos, including dynamically changing focal lengths. We also define several types of camera motions: panning, zooming, orbiting, tracking, etc. and add realistic motion noise to these. This is similar to prior work [27, 56, 58]. We go further, however, to capture real camera motions using hand-held phone and tablet devices as well as an Apple Vision Pro headset for ego-centric camera motion captures. For these captures, we place 3D humans in a 3D scene and users move around the scene to view the virtual subject(s). This induces natural movements with realistic camera shake from both hand-held and ego-centric views. The resulting dataset contains over 27K image sequences with over 8M frames, over 4K diverse body shapes, resulting in 13.3M bounding boxes.

To evaluate the dataset, we train several SOTA HPS regressors using BEDLAM (B1) and BEDLAM2.0 (B2) and evaluate their accuracy. Using B2 results in a significant improvement in accuracy compared with B1 across all standard metrics and the combination of B1 and B2 leads to SOTA performance on human pose estimation in world coordinates.

Similar to BEDLAM, the released dataset includes the videos, ground truth body parameters in SMPL-X format [40], camera motions and focal lengths, clothing assets, 3D hair assets, and depth maps. For assets that we cannot distribute, we provide links to the sources. As with BEDLAM, the assets will enable others to render their own versions of the dataset for specific tasks like egocentric vision. We also provide separate training and testing splits. BEDLAM2.0 is available for research purposes.

2 Related Work

Prior to BEDLAM, numerous synthetic datasets were proposed for training human pose and shape (HPS) regressors [5, 9, 12, 15, 23, 24, 30, 32, 36, 39, 41, 53]; see [6] for a review. Due to limited realism and/or diversity, none of these are able to replace training data extracted from real images. These methods also focus on human pose and not camera motion. While BEDLAM includes a few sequences with moving cameras, the diversity of camera motions and focal lengths is limited and most of the sequences have static cameras.

While early work on HPS estimation focuses on estimating the 3D human pose in camera coordinates, many methods require bodies in world coordinates. Recent methods focus on this problem [27, 38, 45–47, 59, 60, 64] but are limited by the lack of training data with ground truth camera motions and 3D humans together. BEDLAM2.0 is designed to support this direction so, here, we focus on synthetic datasets published since BEDLAM that include varied cameras and camera motions.

Contemporaneous with BEDLAM, SynBody [63] is a synthetic dataset in which each sequence is rendered from 8 static views. While similar to BEDLAM, it is less effective for training HPS regressors [8]. Microsoft's SynthBody dataset [20] uses SMPL-H [43] with a different head compared to SMPL-X and provides only static poses. They demonstrate how it can be used to detect dense 2D keypoints and they use these to fit a 3D body using optimization. They do not use the dataset to train a 3D regressor or show results on standard benchmarks. STAGE [26] takes a different approach, using generative AI to take a 3D body and produce realistic images matching the body pose but with varied visual attributes such as BMI and clothing. They do not use this for training a regressor but, rather, use this to generate benchmarks for evaluation.

PDHuman [58] and BEDLAM-CC [56] tackle the problem of varied focal lengths, which correlate with the depth of the person from the camera; e.g. long focal lengths are used for people far away and short ones for people close to the camera. They generate synthetic training images with widely varied views and focal lengths and show that training on such data improves robustness to real-world cameras. They do not, however, address camera motion.

Recent methods have introduced richer camera motions than those in BEDLAM. For example, EgoGen [28] builds on the BEDLAM assets and re-purposes them for tasks in egocentric vision. They provide a generative process of an agent's motion in a 3D scene. This enables automated collection of video sequences from an egocentric viewpoint. The HumanVid dataset [61] also leverages BEDLAM and adds camera motions using simple rules. They sample multiple random camera locations in

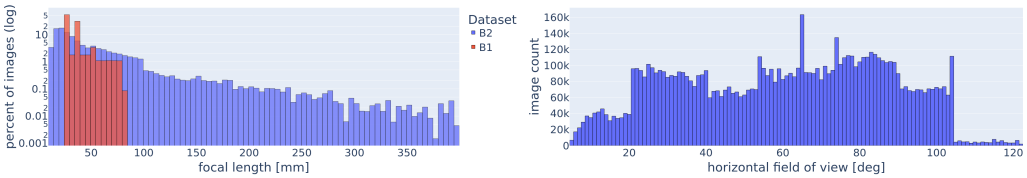

Figure 1: **Camera intrinsics.** (left) Log frequency of focal lengths. Red: BEDLAM; Blue BEDLAM2.0. (right) Histogram of the horizontal field of view (HFOV).

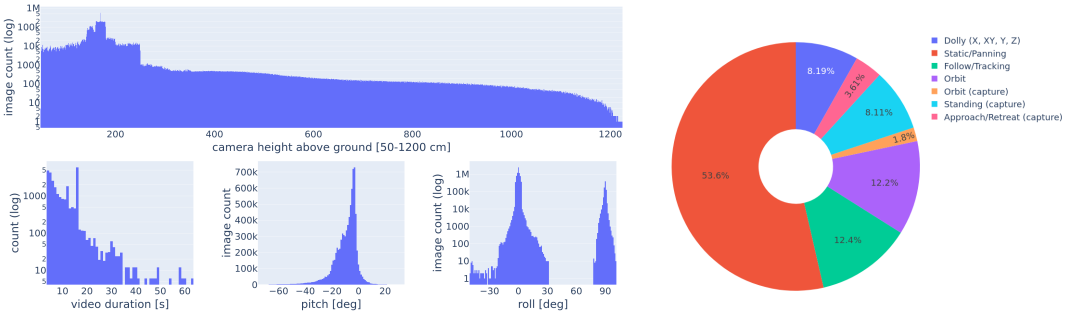

Figure 2: **Camera motion statistics.** (top) Log histogram of camera height above the ground (50-1200cm). (bottom left) Video duration (seconds). (bottom middle) Histogram of pitch. (bottom right) Log histogram of roll. (right) Distribution of different types of camera motion (see text).

space at specific keyframes and point the camera at the subject's face. They then smoothly connect the cameras to create the camera motions. WHAC-A-Mole [65] renders dancing sequences, with pairs of dancers, with varied camera motions including Arc, Pan, Push, Pull, and Tracking shots plus combinations of these. Unfortunately, the synthetic data lacks realism. PACE [27] renders 3D characters in scenes with a moving camera but the amount and diversity of the data is limited. Consequently, they use it only for evaluation of human and camera motion estimates.

3 Dataset: Methods

Here, we describe the key improvements of BEDLAM2.0 that increase its diversity and realism as compared with BEDLAM.

3.1 Cameras

The recent focus of the field is on human pose and shape estimation in scenes with moving cameras. There is limited training data for such scenarios, which makes it hard to train models end-to-end on this task. BEDLAM2.0 addresses this by significantly increasing the complexity and realism of the cameras and their motions.

Focal lengths. BEDLAM intrinsics primarily cover a small Horizontal Field Of View (HFOV) range of $52°$ or $65°$ and are mostly fixed during a shot. Temporal variation of camera extrinsics is very limited since most sequences use a static camera. We address this in BEDLAM2.0 with a much broader focal length coverage. Figure 1 shows the distribution of HFOVs. Specifically, we cover focal lengths from 14mm up to 400mm on a 16:9 DSLR sensor (36 x 20.25mm) that is designed to mimic real-world focal lengths. This is much more diverse than the statistics of Flickr images reported in [38]; see Appendix Fig. 11 for details and a comparison with [38].

Nine percent of all generated videos have varying focal length during the shot by zooming in or out. This is important for realism as it mimics many real-world videos. Start and end focal length values are randomized from a predefined configuration range suitable for the desired location shot and are then keyframed in Unreal Sequencer using Unreal Python automation. Indoor environment shots typically have short focal lengths, whereas long focal lengths are primarily used in outdoor environments.

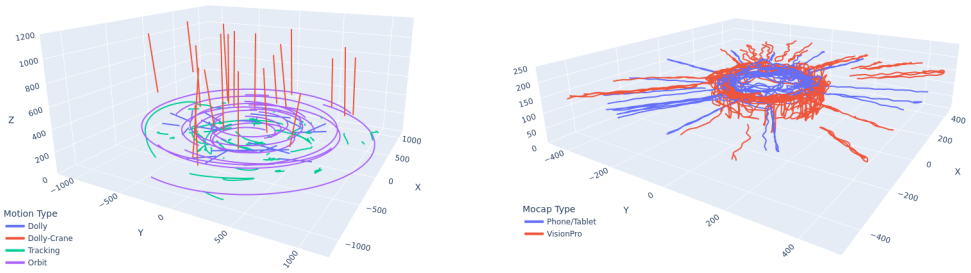

Figure 3: **Sample camera motions used in dataset.** Left: Synthetic, Right: Captured.

Synthetic camera motions. BEDLAM2.0 introduces a variety of auto-generated synthetic camera motions, including static, panning, tracking, dolly, orbit, and zoom, and various combinations of these; see Fig. 3 (left) and the Appendix for details. To maximize variation in camera pose we optionally layer differentiable synthetic Perlin-noise camera shake for location and rotation on top of all these motions. The intensity of this shake effect is randomized. We have options to track individual randomized body parts like the pelvis, spine or head with additional height offset randomization. For shot types that track moving bodies, the changes in extrinsics are keyframed in Unreal Sequencer, similar to the focal length setup. For tracking shots we utilize a custom Unreal Blueprint that uses a camera setup with the Unreal SpringArm component for smooth camera motions via low-pass filtering of the target location. For panning shots we use the default LookAt low-pass rotation filter feature of the Unreal cinematic camera. When low-pass filtering position or rotation we make sure to properly initialize the camera pose to the correct start pose for the first rendered frame before activating the filter. For the majority of shots the camera height above ground is between 0.5m and 2.5m with the exception of crane shots (dolly up/down) where the camera can move up to 12m above ground. This gives top-down views, which are often missing from existing real-world datasets, increasing diversity.

Captured camera motions. To add more realism and diversity, we capture real-world camera motions. These include hand-held camera motions, which we capture with phone and tablet devices as well as egocentric camera motion data captured with an Apple Vision Pro headset. The user views a scene containing 3D humans and we capture three types of camera movements: static location shots of a user standing in same location, orbit shots, and approach/retreat shots. See the Appendix for details. Before rendering, we optionally randomize both handheld and egocentric camera motion data with additional height and distance offsets, pitch offsets, as well as world-space rotation offsets for viewpoint randomization. Figure 3 (right) illustrates a few of these captured motions.

Summary. Together, the synthetic and real camera motions cover a broad range of movement types (Fig. 2 (right)) with a diversity of camera pitches and heights (Fig. 2 (left)). BEDLAM2.0 also contains many longer sequences than B1 (Fig. 2 (bottom left)). In the final dataset, 86.4% of the motions are synthetic, while 13.6% are captured. See the Appendix for further details and `https://b2dash.is.tuebingen.mpg.de/` for detailed statistics.

3.2 Human motions

Our motion pool is composed of 4643 motions in SMPL-X format, as compared to 2311 in BEDLAM. The pool includes diverse motions sampled from AMASS [34], in particular from the datasets CMU [10], KIT [35], BMLmovi [17], BMLrub [52], HDM05 [37], ACCAD [3], Transitions [34], MoSh [33], SOMA [16], PosePrior [4] and DFaust [7]. We go beyond BEDLAM to sample additional motions from the MOYO dataset [51], which contributes complex yoga movements, and the BEAT2 dataset [31], which captures conversational gestures.

Preprocessing. The motions are filtered to exclude actions where the balance of the body depends on an external object, such as sitting, or where the motion is not supported by a ground plane, such as going up and down stairs. Since many mocap motions start and/or end with a T-pose, we automatically identify the frames where the body is in T-pose and exclude these segments from sampling. Furthermore, we subsample the motions to 30 fps and add an offset to the translation parameters to ensure that the body is centered at the origin during the first frame of the motion.

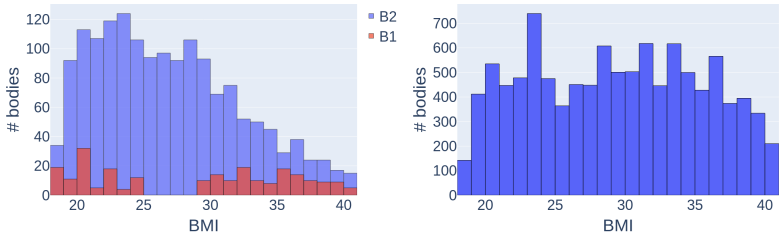

Figure 4: **Body shapes.** Left: Histogram of BMIs for the body shapes in BEDLAM (red) and BEDLAM2.0 (blue). Right: Resampled histogram used to generate more diverse bodies in BEDLAM2.0.

Sampling. From the motion pool, we sample motion segments ranging from a minimum of 4 seconds to a maximum of 16 seconds in length, prioritizing 16-second long segments when available and shortening only when the source motion duration is not long enough. For comparison, the maximum length in BEDLAM is 8 seconds. From this sampling process we obtain a total of 10592 motion segments and 3,231,846 pose frames, of which 1,665,448 (51.53%) appear only once in the dataset.

Retargeting. To augment the animation data with various body shapes, we designed an automated pipeline to retarget the samples from our motion pool to the sampled body shapes. This is necessary because the sampled bodies may have different limb lengths than the original motion-capture subjects. We use the IK-Retargeter tool from Unreal Engine, which retargets a source motion to a target skeleton, minimizing foot-sliding by world-space pelvis location adjustments for the new limb lengths. This significantly increases the shape-motion diversity compared to BEDLAM. Moreover, by adjusting body poses to fit different skeletons, the retargeting increases pose variety. The code is available on the project website.

Hand motion. Due to the complexity of capturing hands with optical motion capture technology, a large fraction of AMASS does not contain hand motion. In order to provide hand pose variation, we augment the data by adding randomly sampled hand motions from the ARCTIC dataset [14] to the existing AMASS motions.

3.3 Body shape

Body shapes in BEDLAM are not as diverse as in the real population. Here we increase the body-shape diversity, particularly for high-BMI bodies. To that end, we sample SMPL-X [40] with 16 shape parameters in accordance with the body mass indices (BMIs) in the CAESAR dataset [42], which includes a diverse range of male and female body shapes. Specifically, we sample 1,615 bodies with BMIs ranging from 18 to 41, ensuring balanced representation across the entire BMI spectrum, as shown in Fig. 4 (left). Note that the BMIs in CAESAR are skewed to BMIs under 30. To provide more shape diversity, we resample this distribution to include more bodies with high BMIs as depicted in Fig. 4 (right).

Note that B1 uses a different version of SMPL-X from B2. For B2, we use the version with a "locked" head, which removes the hair "bun" from the shape space. This is needed for realistic hair groom generation. B1 also uses fewer body shape components (11 vs 16). To enable training and comparison using both datasets, we refit the B1 ground-truth using the B2 16-beta model. The resulting motion files are available for download from the original BEDLAM website (`https://bedlam.is.tuebingen.mpg.de/`).

3.4 Hair

Hair realism and variation in BEDLAM is limited by the card-based hair models used and most subjects do not have hair. Additionally, the hair assets have a license that does not allow redistribution. To address these issues BEDLAM2.0 uses higher quality strand-based 3D hair grooms. This approach models the hair as individual hair strand 3D curves, allowing us to adapt each hair groom to the individual body head shapes. This also improves render quality and results in more accurate hair rendering under HDR image-based lighting conditions with raytraced shadows. Unlike BEDLAM we use hair in all rendered sequences.

Figure 5: **Strand-based hair.** These examples illustrate the realism and diversity, which is much better than in BEDLAM.

a b c d

Figure 6: **SMPL-X Shoes.** (a) SMPL-X foot vs the "sock" foot. (b) Examples from the shoe database. (c) Displacement, normal and texture maps. (d) Shoes rendered on the SMPL-X body.

We contracted a professional VFX studio with previous experience in generating realistic grooms for synthetic data generation [62] to create 40 unique hairstyles for our SMPL-X (AMASS, no head bun) neutral mesh default head shape (see Fig. 5). Each groom has between 50k and 100k 3D strand curves of varying length. Total vertex count for a groom starts at 1 million for straight hair and can go up to 13 million for complex curly hair styles like afro, where each strand contains about 170 vertices. Unlike previous work in this area, we release all of these grooms for non-commercial use. In Unreal, we apply a randomized hair groom to the selected target body and use the hair binding component to adapt it automatically to the target headshape surface. Hair color is determined by a dedicated hair shader, which uses a combination of melanin and redness values to define the 9 hair material presets used for rendering.

3.5 Shoes

BEDLAM, and similar datasets, contain SMPL or SMPL-X bodies with bare feet. This creates a domain gap to real imagery in which people typically wear shoes. This further creates an issue for estimating the body height and foot-ground contact because shoes introduce a displacement between the bottom of the SMPL-X foot and the ground.

Adding shoes to SMPL-X is not trivial since the mesh topology represents the toes. Consequently, as a first step, we smooth out the SMPL-X toes to create a smooth canonical sock-like foot that better matches the shape of shoes (Fig. 6a).

Next, we source shoes from the Google Scanned Objects dataset [13], which contains a wide variety of shoes and textures (see Fig. 6b for examples). We use a subset containing 45 loafers, 6 formal shoes, 9 ballerina flats, 3 flip-flops, 18 boots, 5 football shoes (with traction studded soles), while the remaining 96 are casual sport shoes.

We align all the shoe meshes, scaling them to create a common shoe size corresponding to the neutral SMPL-X mesh. We then align the shoes with the SMPL-X mesh and, for each pair of shoes, bake normal and texture maps to the SMPL-X UV space and compute a displacement map (Fig. 6c) that, when applied to the sock-like foot, deforms its shape to match the shape of the shoe. Subsequently, we add an appropriate upwards translation to the whole body to account for the sole thickness. See the Appendix for further details.

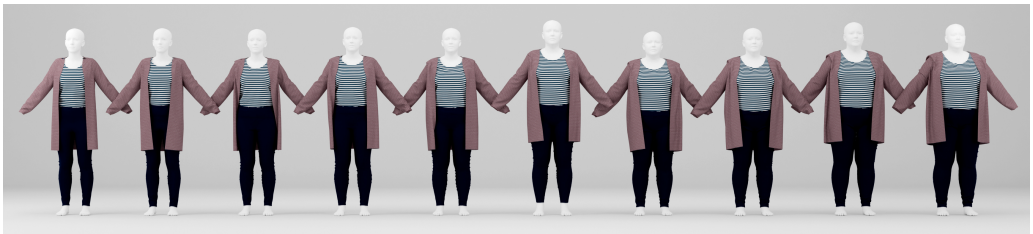

Figure 7: **Graded garments.** Example of a graded outfit draped on bodies of varying BMI.

This approach provides the illusion of a shoe (Fig. 6d) while maintaining compatibility with AMASS motions [34] and leverages the SMPL-X shape space to handle shoe-size changes according to body shape. We only consider flat-soled shoes because heels require a significant change in foot topology and would influence posture and movement; this is future work.

3.6 Clothing

BEDLAM animates bodies in 111 high-quality, detailed, 3D outfits. Unfortunately, the outfits come in only one size and do not fit high-BMI bodies. Consequently, all large bodies in BEDLAM are only rendered with clothing textures on the naked body, limiting realism and making it hard to train methods that estimate body shape under clothing.

We go beyond BEDLAM to add 76 new 3D outfits (187 in total) created by a professional 3D clothing designer using CLO [1]. The dataset contains a wide variety of outfits, from simple dresses to complex multi-layered outfits comprised of multiple garments, such as a man's suit.

We had the designer grade 50 of the outfits into sizes XS, S, M, L, XL, 2XL, 3XL, 4XL, 5XL, and 6XL. We match each outfit size with a reference BMI value, and assign an outfit size to each avatar based on the BMI of the body. This allows us to dress, animate, and render bodies of all sizes in realistic clothing (Fig. 7).

For each outfit, we define a set of texture patterns (from BEDLAM) from which we sample during rendering. Each outfit has a minimum of 6 and a maximum of 28 texture variations, while the median number of texture variations per outfit is 10. As in BEDLAM, cloth simulation is done using CLO and manually checked for quality; see the Appendix for details.

3.7 Scenes and lighting

Similar to BEDLAM we render the animated bodies in many different environments including both simple scenes with high-dynamic-range (HDR) image-based lighting and full 3D environments.

HDRI. We use HDRI environments when the camera is primarily panning and not translating since there is no parallax in this case. We randomly sample from 94 preselected HDR images with varying illumination from PolyHaven [2] and use these as the sole light source to light the bodies. The equirectangular HDR image is re-projected onto a virtual half dome using the Unreal HDRIBackdrop component. To increase the illumination and background variation we randomize the orientation of the rendered subjects and cameras in world space. Raytraced shadows are used to provide realistic dynamic body shadows on the ground plane for HDR images taken in direct sunlight.

3D environments. BEDLAM uses only 5 3D environments, whereas BEDLAM2.0 has 15 high quality, geographically diverse, 3D environments (from Unreal Marketplace and Fab). Only one 3D environment is shared between BEDLAM and BEDLAM2.0 and the other 14 are new. The number of indoor environments increases from 1 in BEDLAM to 9 here. For each environment, we sample between 3 and 25 shot locations depending on the type of environment, creating much more background variation compared to BEDLAM. We also increase lighting variation by using time-of-day randomization for selected environments with properly calculated physically correct sun light direction. We also use different lighting setups like daylight, sunset, overcast and night time when available in the 3D environment.

3.8 Occlusion

The dataset includes a significant amount of occlusion of the bodies, including frame occlusion, self-occlusion, person-person occlusion, and environmental/scene occlusion, not to mention occlusion of the body caused by clothing, hair and shoes.

We performed a quantitative analysis on a randomly selected set of 41.5k images (covering 58.6k rendered bodies) to examine frame, scene, and person-person occlusion. Our results show that 12.7% of the images exhibit more than 20% occlusion, and the top 10% most occluded bodies experience an average of 61.1% occlusion. See Fig. 15 in the Appendix for examples.

3.9 Rendering: Image and Depth data

Data is rendered in Unreal Engine 5.3 on NVIDIA RTX3090/RTX4090 GPUs using the built-in Movie Render Pipeline plugin with deferred rasterizer in cinematic quality settings. Two separate render passes are used for image and depth data generation at a resolution of 1280x720 for a target video framerate of 30 frames per second.

Image data. To achieve realistic motion blur, which is important for work like [11], we use the default 180-degree shutter for motion blur and always render 7 separate temporal subframe images that are combined into the final output image. Rendered image data is saved in lossless compressed PNG and EXR formats. EXR output is used to store the correct camera pose used at render time for each frame in JSON format embedded in the EXR metadata. This approach ensures correct camera pose data when camera shake modifiers are used. We observed that Blueprint-based approaches for camera pose logging like the method used in BEDLAM do not capture these additional shake modifications, which would lead to inaccurate camera ground truth data.

Depth data. We also generate ground-truth depth maps in a separate render pass using center subframe camera pose, resulting in non-blurred depth in EXR format with 16-bit float precision. To ensure correct camera pose consistency between the image and depth pass, we created two custom Unreal Engine C++ plugins. The first plugin ensures that Perlin noise camera shake is deterministic between re-renders with controllable variability. We achieve this by extending the existing camera shake functionality with the option to externally specify the used seed for noise randomization. The second plugin fixes the existing behavior of the Unreal Movie Render Pipeline, which only stores the last subframe camera pose and not the desired center subframe camera pose as ground truth in the EXR metadata. When rendering fast camera motions with 7 temporal samples there are noticeable differences in camera pose between the last subframe and the center subframe. We extend the functionality to log the camera pose for all subframes in EXR metadata with center subframe as the ground truth camera pose reference. See the Appendix for more details.

4 Dataset Statistics

BEDLAM2.0 contains 27480 video image sequences with a resolution of 1280x720 at 30 frames per second resulting in total 8,048,411 PNG images. This results in 12.5M and 862K bounding boxes containing humans with ground truth SMPL-X parameters for training and test set, respectively. These images are generated from 56,338,877 rendered temporal subframes for realistic motion blur in every image of the dataset. The average video length is 10s. We also provide compressed H.264 encoded MP4 videos for all sequences as well as overview images for first, middle and last image of all sequences. For each image we provide world-space ground truth for camera extrinsics and intrinsics and ground truth 3D bodies and their appearance randomization parameters. For 44% of the images we also provide depth images (16-bit float) and a corresponding center subframe render without motion blur in EXR multilayer format. All image renders are organized by camera motion type to facilitate selecting desired camera motion subsets. See the Appendix for a numerical comparison with other datasets. Also please see the project website (`https://bedlam2.is.tuebingen.mpg.de/`) and video (`https://youtu.be/ylyqHnwhpsY`) example sequences illustrating the dataset.

To construct the training and test splits, we held out (i) a subset of 161 body shapes, and (ii) a subset of 597 motions, ensuring that the test set contains exclusively unseen pose and shape parameters; we then render 1824 new sequences in 5 new environments, containing only test bodies and motions, resulting in 449061 test images. Please refer to the Appendix for details on the data split and its usage.

Table 1: Single-frame pose estimation using CameraHMR [38]. See text.

Dataset	3DPW [54]			EMDB [25]			RICH [22]		
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓
BEDLAM1	43.2	68.0	80.7	50.0	88.7	101.6	42.1	75.2	83.2
BEDLAM2	**41.1**	**64.8**	**76.3**	46.5	**74.6**	**86.2**	36.8	70.8	79.4
BEDLAM1+2	**41.0**	65.2	77.7	**46.4**	75.5	87.3	**36.4**	**68.0**	**75.7**

Table 2: World-space evaluation of video-based methods.

	RICH (24)					EMDB (24)				
Models	WA-MPJPE$_{100}$	W-MPJPE$_{100}$	RTE	Jitter	Foot-Sliding	WA-MPJPE$_{100}$	W-MPJPE$_{100}$	RTE	Jitter	Foot-Sliding
DPVO[49] + HMR2.0[18]	184.3	338.3	7.7	255.0	38.7	647.8	2231.4	15.8	537.3	107.6
GLAMR [66]	129.4	236.2	3.8	49.7	18.1	280.8	726.6	11.4	46.3	20.7
TRACE [47]	238.1	925.4	610.4	1578.6	230.7	529.0	1702.3	17.7	2987.6	370.7
SLAHMR [64]	98.1	186.4	28.9	34.3	5.1	326.9	776.1	10.2	31.3	14.5
WHAM [46]	109.9	184.6	4.1	19.7	3.3	135.6	354.8	6.0	22.5	4.4
GVHMR [45]	78.8	126.3	2.4	12.8	3.0	111.0	276.5	2.0	16.7	3.5
PHMR [59]						71.0	216.5	1.4	16.3	3.5
GVHMR [45] - B1	87.3	140.0	2.6	13.5	2.9	112.4	284.6	1.8	17.1	3.5
GVHMR [45] - B2	75.5	120.6	2.4	12.3	2.7	113.7	284.4	2.1	15.9	3.4
GVHMR [45] - B1 + B2	75.8	121.3	2.3	11.3	**2.6**	109.7	273.1	1.7	15.0	3.4
PHMR [59] - B1	85.7	139.4	2.9	12.7	4.0	77.6	211.1	1.4	14.9	3.4
PHMR [59] - B2	75.3	122.4	2.5	11.7	2.8	71.9	197.7	1.4	12.2	3.4
PHMR [59] - B1 + B2	**72.5**	**116.6**	**2.3**	**10.2**	**2.6**	**70.5**	**193.7**	**1.4**	**11.3**	**3.2**

5 Experiments

Since BEDLAM is already widely adopted in the field, we focus on comparing B2 with B1. We use standard metrics for camera-space methods (MPJPE, PA-MPJPE, and PVE); see [6]. For methods that estimate bodies in world space, we use WA-MPJPE$_{100}$, W-MPJPE$_{100}$, RTE, Jitter, and Foot-Sliding; see [27, 46, 64] for definitions. We evaluate on 3DPW [54], RICH [22], and EMDB [25]. These are all real-image datasets with high-quality pseudo ground truth.

Image-based methods. For single-image methods the diversity of the training data (poses, scenes, cameras, etc.) is key to accuracy. While camera motion, per se, is irrelevant for single-image methods, B2 still has a wider range of camera focal lengths and camera poses than B1. We take the current most accurate single-frame method at time of writing, CameraHMR [38], as representative of such methods. As shown in Tab. 1 training on B2 alone produces significantly lower error than training on B1. Training on both B1 and B2 does not provide a clear advantage. Training on B2 also results in a 20% improvement in shape accuracy compared to training on B1 (see Table 3 in the Appendix).

Video-based methods. We use B1 and B2 to train two recent SOTA methods that estimate human motion from video, GVHMR [45] and PromptHMR [59]. For the image-space evaluation of these video-based methods, see the Appendix. Here we focus on the world-space evaluation in Tab. 2. The top half of the table reports the accuracy of other existing methods (from their respective papers). Note that these methods are typically trained using a variety of data including real and synthetic (including B1).

The lower half shows GVHMR and PromptHMR (abbreviated PHMR here) trained on B1, B2 or both. Overall, the combination of B1+B2 offers the best results. This makes sense, since datasets like EMDB contain activities that are present in B1 but not in B2. BEDLAM contains several motions like sitting and climbing stairs for which the rendered videos do not contain supporting objects. Hence, these motions are non-physical given the scene. In BEDLAM2.0, we remove these to focus on physical plausibility in the 3D scene. This actually reduces accuracy on scenarios like stair climbing in EMDB relative to B1. Thus the two datasets are complimentary and users can select whether to include B1 or not, depending on the kinds of motions they anticipate.

What is important to note is that GVHMR and PromptHMR, trained using *only synthetic data*, are more accurate the the originally published versions. Note that the original versions train using B1 together with real sequences; both are improved by adding B2, even when no real data is used. For visualizations and more results, including an evaluation of video-based methods on the B2 test set, see the Appendix.

6 Conclusions and future work

BEDLAM2.0 addresses a key need in the community for an extensive ground truth dataset for training methods to estimate 3D human motion in world coordinates, particularly in sequences with moving cameras and changing focal lengths. This is currently a critical topic for the field. BEDLAM2.0 provides diverse ground-truth camera motions not present in any other dataset while improving on the original BEDLAM dataset in every aspect (body shape, clothing, hair, scenes, shoes). The results with SOTA methods suggest that training on B2 (or B1+B2), with no real data, achieves world-space accuracy exceeding the recent SOTA. We release the rendered video sequences, the ground-truth 3D humans, as well as the 3D clothing, hair and shoes; these are all available at `https://bedlam2.is.tuebingen.mpg.de/`. The only assets that we cannot release are the 3D environments and, for these, we provide a "shopping list" in Table 6 of the Appendix explaining how people can obtain the assets. Code for training, evaluation, rendering are provided, as well as the code we use to retarget AMASS motions to new body shapes. We also provide the trained model checkpoints for CameraHMR, PromptHMR and GVHMR.

6.1 Limitations.

A key limitation of B1 and B2 is that they only support people interacting with the ground and not with other objects. B1 contains motions like sitting or climbing that are not grounded in the 3D scene with object contact. We removed such motions from B2 so that the movements are grounded. Even so, other than foot-ground contact, other body-scene contacts may not be accurate; for example, during a cartwheel, the hands may not properly interact with the ground. Generating realistic synthetic sequences of general human-object and human-human interaction remains an open research problem but is clearly the next step for the field. Such data would support inference of human-object and human-human contact; cf. [65]. As in B1, the motions in B2 are not semantically meaningful in the context of the scene or the motions of other humans in the scene. This limits use of the dataset for some semantic tasks. Like previous models, the bodies do not include children, amputees, or people whose body morphology deviates significantly from the mean (e.g. scoliosis). Similarly, the movements are from healthy individuals, lacking physical impairments, motor disorders, or supportive equipment like canes or walkers. And, of course, there is a still a visual domain gap between B2 and real videos. Despite this, as evidenced by our experiments, B2 is sufficiently realistic to produce SOTA results. An important direction for future work is to add facial motions and audio, which are completely lacking in B1 and B2. This is necessary to develop synthetic data that supports reasoning about direct human-human communication.

6.2 Broader impacts.

Synthetic data of people has a huge advantage over real data, which is typically scraped from the Internet without consent. In contrast, synthetic data reduces privacy concerns. The primary use case of BEDLAM2.0 is to train methods to estimate 3D humans from video; this has positive and negative use cases. We use a custom license that prohibits use of the data for "pornographic, military, or surveillance, purposes," as well as to "create fake, libelous, misleading, or defamatory content."

In addition to supporting work on human motion estimation, BEDLAM2.0 is useful for training and evaluating methods for 3D/4D point tracking, structure from motion estimation with non-rigid motions, depth estimation, optical flow, and dynamic scene reconstruction. Recent work on these topics [57, 67] is limited due to a lack of ground truth sequences for end-to-end training.

Acknowledgments. We thank STUDIO LUPAS GbR for creating the 3D clothing and Meshcapade GmbH for the skin textures. We thank T. Alexiadis, T. Obersat, C. Gallatz, A. Bertler, A. Cseke, A. Kuznetcova, F. Doll, S. Bhor, T. Rakshit, T. Niewiadomsky and V. Fourel for 3D outfit texturing and quality evaluation of the clothing simulations. We thank the Software Workshop at MPI-IS for deploying the dataset-statistics web app.

Disclosure: While MJB is a co-founder and Chief Scientist at Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society.

References

[1] CLO. https://www.clo3d.com/, 2025.

[2] Poly Haven. https://polyhaven.com/hdris, 2025.

[3] Advanced Computing Center for the Arts and Design. ACCAD MoCap Dataset.

[4] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*, June 2015.

[5] 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.

[6] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, June 2023.

[7] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*, 2023.

[9] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3D human recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10533–10545, 2024.

[10] Carnegie Mellon University. CMU MoCap Dataset.

[11] Jerred Chen and Ronald Clark. Image as an IMU: Estimating camera motion from a single motion-blurred image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[12] 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.

[13] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, page 2553–2560. IEEE Press, 2022.

[14] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[15] 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.

[16] Nima Ghorbani and Michael J. Black. SOMA: Solving optical marker-based mocap automatically. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11117–11126, October 2021.

[17] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje. MoVi: A large multi-purpose human motion and video dataset. *PLOS One*, 16(6):e0253157, 2021.

[18] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023.

[19] Artur Grigorev, Bernhard Thomaszewski, Michael J Black, and Otmar Hilliges. HOOD: Hierarchical graphs for generalized modelling of clothing dynamics. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

[20] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look Ma, no markers: Holistic performance capture without the hassle. *ACM Transactions on Graphics (TOG)*, 43(6), 2024.

[21] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[22] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014.

[24] 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.

[25] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *International Conference on Computer Vision (ICCV)*, 2023.

[26] Nikita Kister, István Sárándi, Anna Khoreva, and Gerard Pons-Moll. Are pose estimators ready for the open world? STAGE: Synthetic data generation toolkit for auditing 3D human pose estimators, 2024.

[27] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. PACE: Human and camera motion estimation from in-the-wild videos. In *International Conference on 3D Vision (3DV 2024)*, March 2024.

[28] Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, Yan Zhang, Marc Pollefeys, and Siyu Tang. EgoGen: An egocentric synthetic data generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14497–14509, June 2024.

[29] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Mannequinchallenge: Learning the depths of moving people by watching frozen people. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4229–4241, December 2021.

[30] 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.

[31] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1144–1154, June 2024.

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

[33] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, November 2014.

[34] 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*, pages 5442–5451, October 2019.

[35] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The KIT whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336, 2015.

[36] 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.

[37] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[38] Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. In *International Conference on 3D Vision (3DV)*, 2025.

[39] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[40] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[41] 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.

[42] 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.

[43] 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.

[44] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild, September 2020.

[45] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia*, 2024.

[46] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2070–2080, June 2024.

[47] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[48] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021.

[49] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36, 2024.

[50] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xR-EgoPose: Egocentric 3D human pose from an hmd camera. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7727–7737, 2019.

[51] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[52] Nikolaus F. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2(5):2–2, September 2002.

[53] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017.

[54] 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.

[55] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with FisheyeViT and diffusion-based motion refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 777–787, 2024.

[56] Shengze Wang, Jiefeng Li, Tianye Li, Ye Yuan, Henry Fuchs, Shalini De Mello, Koki Nagano, and Michael Stengel. BLADE: Single-view Body Mesh Learning through Accurate Depth Estimation. In *IEEE/CVF Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[57] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[58] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV)*, 2023.

[59] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *IEEE/CVF Conf. on Comp. Vision and Pattern Recog. (CVPR)*, June 2025.

[60] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *European Conference on Computer Vision*, 2024.

[61] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. HumanVid: Demystifying training data for camera-controllable human image animation. In *NeurIPS Datsets and Benchmarks Track*, 2024.

[62] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021.

[63] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. SynBody: Synthetic dataset with layered human models for 3D human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20282–20292, October 2023.

[64] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023.

[65] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, Lei Yang, and Ziwei Liu. WHAC: world-grounded humans and cameras. In *European Conference on Computer Vision*, pages 20–37. Springer, 2024.

[66] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[67] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2025.

7 Appendix

In this Appendix, we provide additional details in roughly the order they are mentioned in the main paper. Please see `https://b2dash.is.tuebingen.mpg.de/` for a wide range of dataset statistics, presented graphically.

Note: We maintain a list of known issues on the project website at `https://bedlam2.is.tuebingen.mpg.de/`. If you identify a problem with any part of the dataset please send a report to `bedlam@tue.mpg.de` and include a "[BEDLAM2]" tag in the subject line to indicate that your report is specific to BEDLAM2.0.

7.1 Details of the camera motions

Synthetic camera motion generation. We extend the limited camera motions and focal lengths of BEDLAM by adding a new camera movement component to the render pipeline pre-processing stage. It uses randomizable camera movement configurations and outputs the desired movement definitions as camera pose keyframes for the Unreal Sequencer in JSON format. This is then used with Unreal Python to fully automate the generation of needed Unreal Level Sequences for rendering.

Synthetic camera shot types:

- Panning shots at static location, optionally augmented by tracking of target body parts
- Tracking camera shots for a moving target body. Maintains distance and optionally also viewpoint to the target body. Target location and rotations are low-pass filtered to ensure smooth camera motions.
- Dolly shots (left-right, forward-backward, diagonal, up-down/crane)
- Orbit shots (tracking fixed target location or moving target body)
- Zoom shots with varying focal length, can be combined with other shot types
- Camera shake with Perlin noise with randomized intensity, can be combined with other shot types

Captured camera motions. All motion capture shots are captured for a target stimulus at the origin which allows us to later randomize the viewpoint onto the subject by rotating the mocap data. We also vary the strength of the capture device shake between the various capture sessions.

Handheld footage is captured in landscape or portrait mode with Apple iPhone Pro 14 and Google Pixel 4a phone devices and an iPad Pro 11 tablet device running the Unreal VCam application. This application captures camera extrinsics with the ARKit/ARCore tracking components which estimate 6DOF device pose through visual odometry by sensor fusion of device camera and IMU data, similar to device pose capture approaches in [11]. The camera pose is sent from the handheld device to a Windows PC running a custom Unreal Engine scene with target stimulus. A real-time render of the scene is streamed back from the PC to the handheld device to provide immediate feedback on the current camera pose values, helping the operator to properly frame the stimulus. The PC is also recording the camera pose which we later auto-export from its Unreal-specific binary format into a reusable JSON representation compatible with the synthetic camera motion generation pipeline.

Egocentric footage is captured on Apple Vision Pro mixed-reality headset running a custom Unity 6 3D head pose capture application. It provides an egocentric 90fps real-time render of the stimulus scene based on the head pose of the user. We record the user head pose at 90Hz. To ensure consistent data capture timestamps we first record head pose into a large pre-allocated data array in memory and only save to local device storage at the end of the capture session. Captured data can optionally be played back on device and visualized with a virtual camera 3D model for initial quality assessment. This capture setup is completely self-contained without external device dependencies and can be used wherever there is enough walking space for the desired camera motion. This approach allows us to capture camera motions that go beyond the space limitation of typical optical motion capture studios.

7.2 Dataset statistics

BEDLAM2.0 includes:

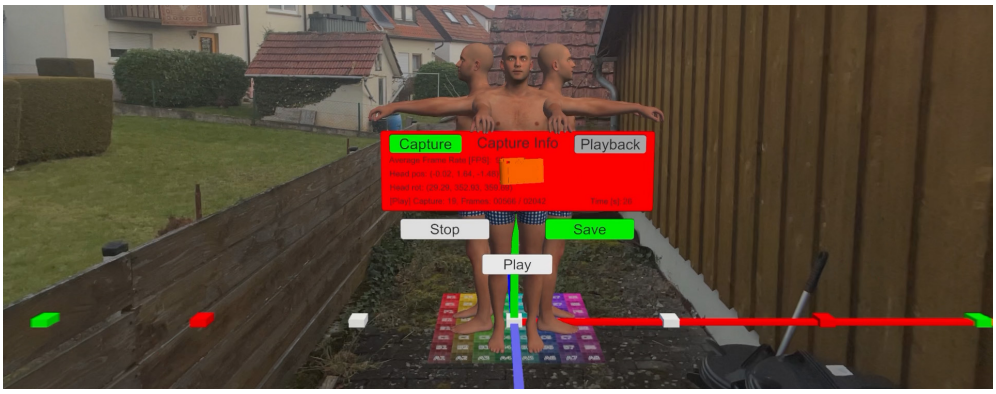

Figure 8: **Egocentric Camera Motion Capture** In-the-wild 90Hz camera pose capture setup on Apple Vision Pro, where our custom Unity 6 app renders a real-time 3D reference stimulus. The user can start, stop and review captures by interacting with the provided 3D user interface.

- 8M images with realistic motion blur using 7 temporal subframes
- 27480 video sequences at 1280x720 resolution, 30fps
- 74.52 hours of video
- ground truth SMPL-X bodies with 16 shape coefficients
- 1,615 diverse body shapes including high and low BMI
- 4,643 motions sampled from AMASS, MOYO and BEAT2
- 187 unique clothing outfits
- 182 unique shoes
- 40 unique hairstyles with 50-100K 3D strands
- 10,592 unique combinations of body shape, motion, and clothing
- widely varied focal lengths and camera motions with realistic noise and ground truth
- 94 HDRI environments and 15 detailed 3D environments
- ground truth depth maps
- 26TB of data

Note that the core dataset distribution includes the images together with the associated SMPL-X ground truth and camera ground truth. We also provide the 3D clothing, shoes, and hair; Subsection 7.14 describes the assets in detail and how to obtain them.

Detailed interactive plots with comprehensive statistics that cover the rendered images, cameras and camera motions as well as bodies and animations can be found at this website: `https://b2dash.is.tuebingen.mpg.de/`

7.3 Shoes

In contrast to previous datasets based on SMPL and SMPL-X, here we add shoes to the model. Roughly 40% of the rendered images contain bodies with shoes. It is easy for users of the dataset to sample sequences with our without shoes as needed.

Shoes are represented using the SMPL-X UV space by defining texture, displacement, and normal maps, which we obtain from the Google Scanned Objects dataset. Here we provide details of how we transform shoes from the dataset and apply them to SMPL-X.

Stocking feet. We first modify the foot of the SMPL-X template mesh to remove the shape of the toes, while keeping the mesh topology the same (see Fig. 6a in the main paper). We find that applying the standard shape parameters (β) to this modified template mesh produces natural looking variations in foot shape. We make this modified SMPL-X template, with stocking feet, available for research.

Figure 9: **New outfits.** We added 76 new and more complex clothing outfits to BEDLAM2.0.

Texture Extraction. We align, rotate and scale the raw scans from the dataset to create a common shoe size that corresponds to the neutral SMPL-X base foot (zero-pose, zero-shape). We then translate all shoes so that they align with the SMPL-X base feet.

Textures are extracted from the shoe models using the xNormal software (`https://xnormal.net`). It works by raycasting from the right foot to the right shoe to get color, normal and displacement values. These values are mirrored for the left foot.

We make the code for putting shoes on SMPL-X available and this contains all the details.

Sole thickness and ground contact. Sole thickness for each shoe is calculated as the mean displacement value in the displacement map corresponding to the sole of that shoe. The body is then translated vertically by the mean amount.

7.4 Hair

Note that we cannot say that having realistic hair actually improves HPS accuracy as we have not performed an ablation study that removes hair. While this remains unclear, realistic hair may serve other uses like training generative models of 3D hair and hair motion. We hope that the 3D assets will find uses beyond our present application.

7.5 Clothing

More examples of new complex outfits and graded outfits are shown in Figure 9 and Figure 10 respectively, and in the supplemental video (`https://youtu.be/ylyqHnwhpsY`). Note that an "outfit" may include multiple pieces of clothing; e.g. a man's suit includes the pants, dress shirt, and jacket. These individual pieces can be rendered individually or in new combinations. This gives users of the dataset flexibility in generating new data.

Simulation. The garments in motion are physically simulated using CLO. We pin some of the vertices of the outfit mesh to the body to avoid garments sliding and falling. Clothing simulation is extremely sensitive to mesh interpenetration; garments can get stuck to body parts and be pulled and deformed in extreme and unnatural ways, causing the simulation to break. Due to the fact that the SMPL-X model does not have the ability to deform the body based on contact, body self-interpenetration is quite common. Additionally, when we apply 3D motions captured on a slim subject to a high-BMI body shape, this can cause self-penetration of the body parts. In order to reduce the number of simulation failures, we temporarily remove the hands of the body mesh during simulation because they are very likely to interpenetrate the body, especially in the thigh region. A correct outfit size assignment (described in Section 3.6 of the main paper) is also crucial to reducing the number of failures, allowing us to successfully simulate people with a wide variety of body shapes and complex outfits. The simulation results are visually checked and rated in order to exclude failures.

Figure 10: **Grading.** More examples of graded outfits. Here are shown, from left to right, sizes XS, M, L, 2XL, 3XL, 5XL and 6XL.

Due to their complexity of the yoga motions in the MOYO dataset, we do not use clothing simulation for these. Instead we use texture maps, which look like tight-fitting yoga clothing. This type of clothing is appropriate for these motions.

7.6 Depth data

Since we are rendering the dataset, we can also render depth maps. While we do not use these for training HPS methods, there are many applications where it is useful to have the depth maps associated with the video sequences. We render approximately 44% of the images with the associated depth map; see `https://b2dash.is.tuebingen.mpg.de/`.

Note that, to create realistic motion blur, each RGB frame of the video is generated from 7 rendered subframes. Note also that the depth data is not blurred. Thus it is critical that the depth data correspond to a specific point in time, which we chose to the center subframe in the set of 7.

By default, Unreal saves the last subframe camera information and not the center one in the generated EXR files. To address this we extended the existing MovieRenderQueue C++ plugin with functionality to store camera ground-truth values for every subframe. This allows us to then properly store the ground truth for the center subframe in the EXR image metadata section.

Data generated before 10/2024 used the last subframe and we fixed this with post processing. Specifically, we first determined the fixed time delta between the last and center subframe by modifying the Unreal C++ source code to obtain this information. For our use case with 7 temporal samples at 30fps render rate, we obtained a subframe delta time of 0.002375s which results in 0.007125s time difference between the last subframe and the center subframe when using 7 temporal samples. We then used that information to resample the camera pose ground truth from the last subframe to the center subframe using Piecewise Cubic Hermite Interpolating Polynomials for a tight fit to original data without overshooting. This allowed us to re-render the depth at the center subframe with high accuracy. We also log this information in the camera ground truth JSON files so that it is clear if the camera pose data was resampled in post or correctly obtained from our custom center subframe ground truth render plugin.

7.7 Normal data

BEDLAM2.0 does not contain normal image data. To help researchers who are interested in this type of data we are releasing the BEDLAM2.0 render pipeline code for Unreal Engine 5.3, similar to our previous BEDLAM render code release for Unreal Engine 5.0. Access details are provided on our project website (`https://bedlam2.is.tuebingen.mpg.de/`). The new 5.3 code includes the functionality to also output camera-space or world-space normal images in EXR format. This functionality can be useful for downstream tasks like normal estimation or avatar creation from a single image.

7.8 Comparisons with other datasets

It is hard to fully compare synthetic datasets numerically since they rarely report all the details and the details that are reported often differ. Here we try to bring together the relevant, and known, details to enable numerical comparison.

There have been several static image datasets like SynthBody [20], BEDLAM-CC [56], and PDHuman [58] that are all similar to BEDLAM but address different issues. SynthBody adds more facial detail than is present in BEDLAM, while the other two datasets focus on close-up shots of the body with wide-angle lenses. We focus on the datasets that introduce camera motion.

SynBody [63]: This dataset is very similar to BEDLAM in terms of goals and approach. They sample 10K body shapes from the SMPL-X shape space and generate 6,960 sequences with 1.2M images and ground truth SMPL-X. These are generated in 6 3D scenes. They use 1,187 motions from AMASS. The dataset is generated from 4 fixed camera viewpoints; it does not include camera motion. As reported in [8], SynBody is not as effective for training as BEDLAM. The reasons are not completely clear but it likely has to do with the visual quality, which is lower for SynBody.

WHAC-A-Mole [65]: This dataset is built on SynBody and inherits its limited realism. What it adds, however, are sequences with human-human interactions, including dancing, together with diverse camera motions. Like BEDLAM2.0, they use a range of standard camera motions but, in addition, they use combinations of these. Overall, the dataset has 1.46M crops and 2434 sequences.

PACE [27]: Camera trajectories use heuristics and include dolly zoom, arc motions, tracking shots, and motions from the MannequinChallenge dataset [29]. The bodies are scans from RenderPeople (see [39]), the motions are from AMASS [34], and the 3D scenes are from Unreal Marketplace. The dataset contains only 25 video sequences with 1-8 people per sequence.

HumanVid [61]: HumanVid generates synthetic sequences with either SMPL-X bodies or anime characters; we focus on the former here. For this, they use BEDLAM bodies, motions and clothing. They appear to use only one body per sequence. Where they go beyond BEDLAM is in designing a rule-based camera motion generation pipeline. Given a body, they sample camera locations in a semicircle in front of the body, point the camera at the person, and randomly sample the camera roll. They form a camera path between these camera keyframes using an interpolating spline. With this, they generate 50K clips with a total of 8M frames.

EgoGen [28]: Goes further to use generated human motions in varied environments, effectively enabling the capture of an infinite amount of synthetic egocentric video. There have been several prior efforts [50, 55] to create egocentric data but EgoGen is more realistic in that it is built on BEDLAM assets. Rather than use physics simulation for the clothing as done here, they use a neural garment simulation method [19] to make the process more automatic and scalable. They generate two datasets. The depth dataset has 105,000 depth images, while the RGB dataset has 301,073 images. The key property of this dataset is that the generated motion of the body drives the camera, producing diverse and natural looking camera motions. We get similar motions for BEDLAM2.0 by directly capturing them from an Apple Vision Pro or handheld device.

7.8.1 Focal Length Distribution Comparison

In addition to synthetic datasets, we compare the distribution of camera intrinsics in BEDLAM2.0 against real-world images from the Flickr HumanFOV dataset introduced in CameraHR [38]. To minimize the domain gap between synthetic and real world cameras, BEDLAM2.0 is designed to cover the space of FOVs observed in the Flickr dataset (see Fig. 11). While the Flickr HFOV (Horizontal Field of View) values are derived from real EXIF metadata and capture natural camera usage patterns, BEDLAM2.0 samples HFOV values synthetically.

Note that the Flickr data has distinct "spikes" for common focal lengths. BEDLAM2.0 is quite different in that it is a video dataset and many sequences include changing focal lengths, i.e. zooming during the camera motion. This results in a more even distribution of HFOVs.

7.9 Human motions and bodies

To construct our pool of moving bodies, we repeatedly sample bodies from a set of 1,615 body shapes and assign them a motion from our set of 4,643 motions, whose composition is shown in Figure 12.

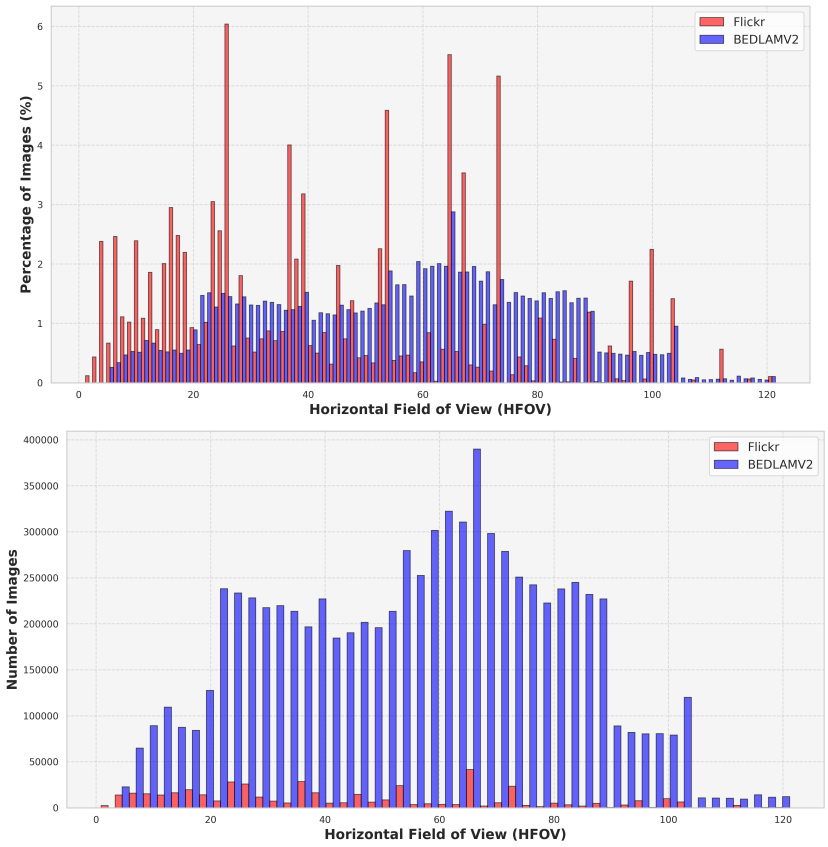

Figure 11: HFOV comparison between BEDLAM2.0 and real Flickr data from [38]. (top percentage of HFOVs. (bottom) absolute number of HFOVs.

As mentioned in subsection 7.5, we do not run clothing simulation for the 171 motions in the MOYO dataset. We obtain a total of 10,592 motions with clothing simulation that are used multiple times across different render sequences with varied clothing textures.

We use the SMPL-X neutral body with the locked head – no hair bun and 16 β shape parameters. This differs from the version in BEDLAM. The original BEDLAM uses a version of SMPL-X without the locked head, which can produce a hair "bun". This bun produces a bump at the back of the head with female body shapes, and this makes it difficult to simulate strand-based hair realistically; for hair simulation, we need a proper scalp shape. Moreover, the bodies in BEDLAM are represented with only 10 β shape parameters.

We provide a version of the original BEDLAM ground truth in the format of B2, making it easy for people to train HPS models using a combination of B1 and B2 data.

Train-test split. We reserve a holdout test set of body shapes and motions. The test set is composed by 161 body shapes (10% of the total) and 597 motions (13% of the total). Note that these bodies and motions may appear in the training set images, but their ground truth SMPL-X parameters are not provided. This prevents people from training on these characters.

Note also that the motions come from AMASS so, theoretically, people could exploit this to gain an advantage in accuracy. However, since we retarget these motions to new body shapes, they are not identical to the AMASS motions.

The BMI distributions for the training and test bodies can be seen in Figure 13, while the numbers of motions sampled from each dataset is shown in Figure 14. The number of unique motion sequences with clothing simulation that are part of the test set amounts to 641 (6.61% of the total); these are

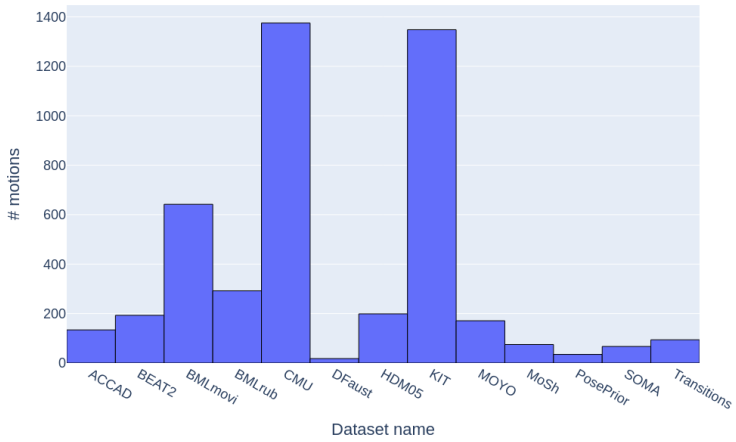

Figure 12: **Motion pool** Number of motions sampled from each dataset, including BEAT2, MOYO and a subset of the datasets in AMASS.

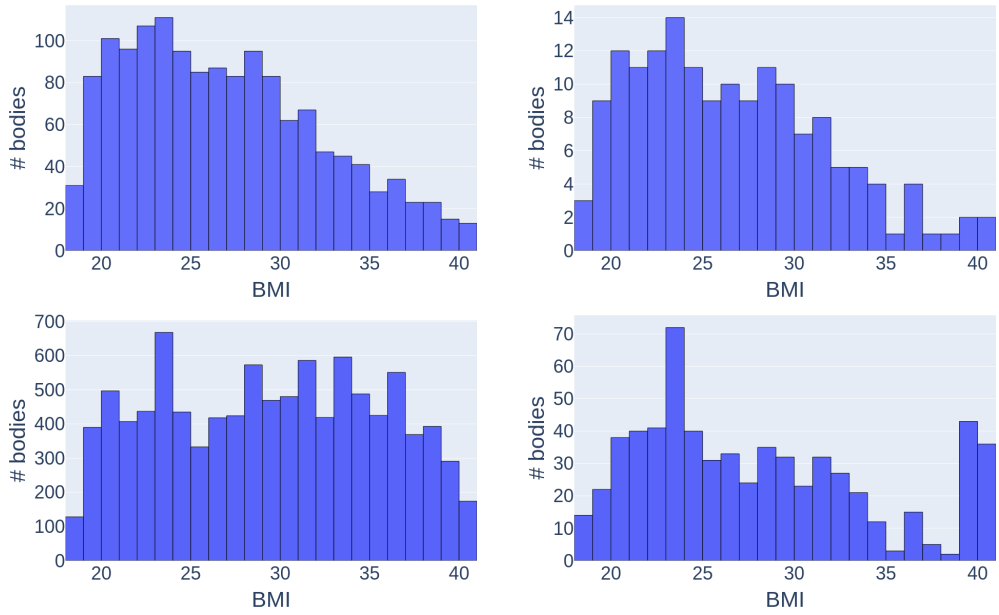

Figure 13: **Train-test body split.** Left: BMI distribution of body samples in the training set before (top) and after (bottom) oversampling high-BMI examples to balance the dataset. Right: Corresponding BMI distribution in the test set, shown for comparison.

used to generate 1,824 new render sequences across five novel environments, featuring only test bodies and motions, resulting in a total of 449,061 test images.

7.10 Body occlusion

Our dataset covers the following common occlusion phenomena frequently observed in real-world scenarios: frame occlusion, scene occlusion, self occlusion, person-person occlusion and occlusion caused by 3D clothing. See Figure 15 for some examples.

7.11 Visualizing the dataset

Figure 16 shows a few example frames from the dataset together with ground truth bodies.

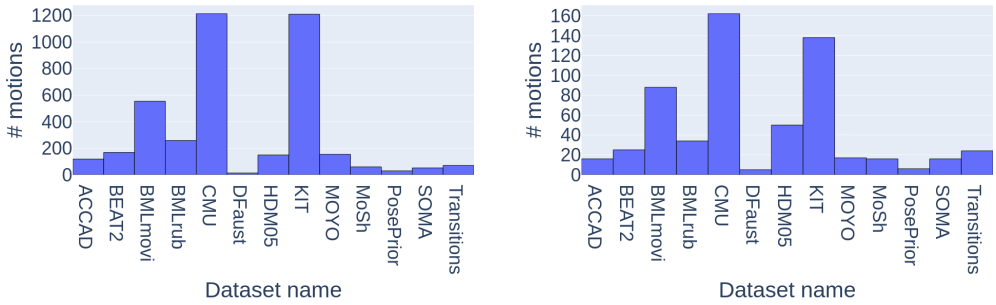

Figure 14: **Train-test motion split.** Left: number of motions from each motion dataset in the training set. Right: Corresponding numbers of motions sampled for the test set.

Figure 15: **Body occlusion examples.** Top row: camera frame occlusion, top right: self occlusion, bottom left: scene occlusion, bottom right: person-person occlusion. All images also show occlusion caused by clothing.

The **Supplemental Video** provides a quick overview of all the dataset components and camera motions: `https://youtu.be/ylyqHnwhpsY`.

Besides PNG image sequences, we also provide all 27480 dataset video sequences in much smaller MP4 video format on our project website (`https://bedlam2.is.tuebingen.mpg.de/`). Small dataset samples for various camera motion types are also available for download.

7.12 Experiments

Shape Evaluation of image-based methods. We train the image-based human pose and shape estimation method CameraHMR [38] on the B1, B2, and B1+B2 datasets. Since standard image-based human pose and shape (HPS) benchmarks exhibit limited shape variability, we use the B2 test set to evaluate shape accuracy. To isolate shape error from pose error, we report results using the PVE-T-SC metric [44], which computes the Per-Vertex Error (PVE) after bringing the predicted and ground-truth meshes to a T-pose and performing scale alignment. As shown in Table 3, training on BEDLAM2.0 reduces the PVE-T-SC error by approximately 20%, demonstrating a substantial improvement in shape estimation accuracy.

Training of video-based methods. We train GVHMR [45] and PromptHMR [59] using only the AMASS [34], B1, and B2 datasets during the video training phase. Both models are trained for 500

Figure 16: **Ground truth bodies.** Example frames from the dataset along with the ground truth bodies projected into the frame.

Table 3: Shape accuracy comparison using PVE-T-SC.

Model	PVE-T-SC ↓	Improvement (%)
BEDLAM1	8.85	–
BEDLAM2	**7.20**	18.6
BEDLAM1+2	7.44	15.9

epochs, and we report results from the final checkpoint. To improve global motion stability, we apply foot skating post-processing to both methods, which leverages foot contact predictions to reduce unrealistic sliding artifacts.

Image-space evaluation of video-based methods. We use B1 and B2 to train two of the recent SOTA methods that estimate human motion from video, GVHMR [45] and PromptHMR [59]. Table 4 shows the results using the camera-space metrics (PA-MPJPE, MPJPE, and PVE). These errors are lower than for the single-image method except for PA-MPJPE on RICH. B2 or a combination B1 and B2 provides the best results for all but the MPJPE and PVE on the RICH dataset where B1 is best.

Evaluation of video-based methods on BEDLAM2.0 test set. We evaluate GVHMR[45] and PromptHMR [59] on the B2 test set. The world coordinate metrics show that B2 is more challenging than existing benchmarks like EMDB [25] and RICH [22].

See the main paper, Table 2, for the evaluation of the video-based methods on real video benchmarks.

A key property of the B2 test set is varying focal lengths within sequences; e.g. dolly zoom. This creates difficulties for SLAM-based approaches, which typically assume static camera intrinsics. PromptHMR [59] relies on a metric SLAM method [21, 48] to transform human motion from camera to world coordinates. Therefore, PromptHMR fails more often than GVHMR on sequences with focal length variations. GVHMR is more robust because it only uses the angular velocity of camera motion, making it less sensitive to focal length changes. This highlights that BEDLAM2.0 is sufficiently challenging to drive the field to develop robust methods that cope with natural camera movements.

Table 4: Camera-space evaluation of video-based methods.

Dataset	3DPW [54]			EMDB [25]			RICH [22]		
	PA-MPJPE ↓	MPJPE ↓	PVE↓	PA-MPJPE↓	MPJPE↓	PVE↓	PA-MPJPE ↓	MPJPE↓	PVE↓
GVHMR [45] - B1	39.8	60.7	73.9	45.5	75.3	87.7	42.0	70.1	78.7
GVHMR [45] - B2	38.4	58.6	70.4	44.0	72.2	84.1	37.9	67.4	76.5
GVHMR [45] - B1+B2	37.6	**57.2**	70.4	44.0	74.8	87.0	**37.2**	65.3	74.3
PHMR [59] - B1	38.2	57.5	69.2	42.1	80.5	92.7	38.5	**62.1**	**70.4**
PHMR [59] - B2	**37.0**	57.3	**67.5**	**40.2**	73.5	84.2	37.8	69.0	78.4
PHMR [59] - B1+B2	37.2	57.6	68.3	**40.2**	**69.2**	**80.2**	37.3	66.2	75.2

Table 5: World and camera space evaluation of video-based methods on B2 test set.

	BEDLAM2							
Models	PA-MPJPE	MPJPE	PVE	WA-MPJPE	W-MPJPE	RTE	Jitter	Foot-Sliding
GVHMR [45] - B1	57.2	98.6	83.3	240.6	502.2	6.3	16.1	3.0
GVHMR [45] - B2	38.8	66.1	55.7	203.9	444.1	5.3	16.0	2.5
GVHMR [45] - B1 + B2	36.4	62.0	52.0	**195.3**	**440.3**	**4.8**	**14.1**	**2.4**
PHMR [59] - B1	35.1	63.0	54.7	220.8	781.6	6.1	16.1	3.0
PHMR [59] - B2	33.5	57.2	48.6	230.3	801.5	6.4	16.4	3.3
PHMR [59] - B1 + B2	**32.0**	**55.7**	**47.5**	223.2	781.2	6.0	16.0	3.0

7.13 Computational costs

Rendering. To render the entire training dataset, it took approximately 467 hours (383 hours for the image (PNG) files and 84 hours for the depth maps). We initially rendered with an NVIDIA RTX 3090 GPU (60% of generated PNG images) and later switched to a new PC with an RTX 4090 GPU (40% of generated PNG images). Consequently, the overall time needed to re-render the dataset on a RTX 4090 GPU would be less than reported here. The observed GPU utilization varied from 30% to 100%, with HDRI renders in combination with complex hair grooms having the highest GPU utilization and benefiting the most from the RTX 4090 upgrade.

Clothing simulation. The clothing simulations were performed using CLO fashion design software on machines equipped with Intel Xeon CPUs operating at frequencies between 2.0 and 2.5 GHz. Each simulation utilized 12 CPU cores, as using a higher number of cores was found to degrade performance. Under this configuration, simulations—ranging from 120 to 480 frames in length—required an average of 0.8 hours to complete. A total of 10,592 simulations were generated for the training and test datasets, resulting in approximately 8,579 CPU hours on 12-core, 2.0–2.5 GHz processors.

7.14 Assets

Table 6 describes all third-party assets used in making the dataset. Note that some of the 3D environment assets may have a "no-GenAI" restriction. BEDLAM2.0 does not use GenAI in its creation and is designed to further research on human pose and shape estimation. We imagine that there will be other uses (e.g. generative ones) of the dataset beyond what we designed it for. Users are responsible for ensuring that their use case aligns with the asset licensees.

Table 6: Third-party assets used for rendering BEDLAM2.0. Most 3D environments were purchased from Unreal Marketplace and its successor fab.com. Evermotion indoor assets were purchased directly from the vendor as a bundle. Please check individual vendor licenses for further details on Generative AI usage permissions.

Asset Type	Name	Source
Body Textures	Bald Head Versions	Meshcapade GmbH, https://meshcapade.com, CC BY-NC 4.0
Clothing Textures	B1 WowPatterns	BEDLAM, https://bedlam.is.tuebingen.mpg.de/
Environment - HDRI	Various HDRIs	Poly Haven, https://polyhaven.com/hdris, CC0 Public Domain
Environment - 3D	ai0805	Archinteriors for UE, https://evermotion.org/
Environment - 3D	ai0901	Archinteriors for UE, https://evermotion.org/
Environment - 3D	ai1004	Archinteriors for UE, https://evermotion.org/
Environment - 3D	ai1101	Archinteriors for UE, https://evermotion.org/
Environment - 3D	ai1102	Archinteriors for UE, https://evermotion.org/
Environment - 3D	ai1105	Archinteriors for UE, https://evermotion.org/
Environment - 3D	archmodelsvol8	https://www.fab.com/listings/910a05ca-4f7a-4aac-9c1b-c0bf7aabfbd8
Environment - 3D	busstation	https://www.fab.com/listings/55c97991-d732-4f63-a831-d38843fb5fb0
Environment - 3D	chemicalplant	https://www.fab.com/listings/a70632d1-f2d2-4b4d-a621-0dc5c3b259fd
Environment - 3D	citysample	https://www.fab.com/listings/4898e707-7855-404b-af0e-a505ee690e68
Environment - 3D	middleeast	https://www.fab.com/listings/b46926e1-fe3c-4c20-83f0-8be8ee5e8de5
Environment - 3D	rome	https://www.fab.com/listings/12ccee26-1515-4ae9-80a2-cd6402346447
Environment - 3D	stadium	https://www.fab.com/listings/d7cfc283-bf41-46a0-b7cd-789476c3263d
Environment - 3D	yakohama	https://www.fab.com/listings/b46926e1-fe3c-4c20-83f0-8be8ee5e8de5
Environment - 3D	yogastudio	https://www.fab.com/listings/ba2d42c5-c9de-49f4-a681-9d0c8970a670

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the dataset and its contents.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: These are discussed in the conclusions and future work section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This dataset paper does not make a theoretical contribution.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The dataset is made available for research purposes and all assets that we control will be made available so others can replicate the work. For assets from third parties that prohibit redistribution, we provide links so people can obtain them. This was a key principle of the BEDLAM dataset and several groups have used the BEDLAM assets to render their own specialized datasets, e.g for egocentric vision. With BEDLAM, the hair was the only body-related asset that could not be shared for license reasons. So for BEDLAM2.0, we hired a graphics art firm to create high-quality strand-based hair grooms and we can share these.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The dataset is submitted with the paper since this paper is in the dataset and benchmarks track. The dataset is usable as released. We will release more support code when the dataset is made public to make it easier to use, e.g. for training HPS regressors. As uses evolve, so will our support code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details of how the dataset is constructed are provided in main paper and the Supplementary Material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: For the dataset itself, which is our core contribution, the question does not apply. For the evaluation of the dataset, we follow the standard methodology of the field which, sadly, does not include statistical significance testing. The tables of results comparing

with prior methods often rely on the results reported in papers and it is impossible for us to compute variances since these were not reported.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the details of the software and hardware used to create the dataset. We do not provide details for the HPS methods we evaluate the dataset on since these are not our contributions. We make no claims about these previously published methods.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: While the dataset contains people, it is purely synthetic with no real people; there are no privacy issues. The textures used span the gamut of human skin tones. We explicitly include a diverse range of body shapes including both high and low BMI bodies. We have been careful to only use assets to which we have the rights.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper makes available a synthetic dataset of humans in motion. The synthetic nature of the dataset means that it does not violate privacy. This is a huge advantage over most real-video datasets, many of which are scraped from the internet without permission. The dataset should facilitate research on topics related to 3D humans by obviating, or reducing, the need for images/videos of real people.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Because of the nature of the data (people), we release it with a custom license that prohibits its use for pornographic, military, or surveillance, purposes. It also prohibits the use to to create fake, libelous, misleading, or defamatory content of any kind excluding analyses in peer-reviewed scientific research.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We mention and respect licenses, as evidenced in our previous BEDLAM dataset release. All 3D Unreal scenes are purchased from commercial vendors. We list all the purchase links in SupMat for users who want to recreate our data. We release only the parts of the dataset where we are the copyright owner under our custom BEDLAM2.0 license, free for non-commercial use and similar to BEDLAM. Our shoe representation is derivative work and will be released under the same license as the source data license (CC BY 4.0). We are the owner of the hair grooms, clothing garments, and body shapes. The motions from the AMASS dataset are all available for research with individual licenses for each subset of AMASS (all used subsets are cited here). The skin textures are provided by Meshcapade GmbH under CC BY-NC 4.0 for research purposes.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The details of the assets are provided in the Supplemental Material and in the download area of the project website (`https://bedlam2.is.tuebingen.mpg.de/`).

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: We do not use crowdsourcing and this is not human subjects research. We do not collect any data from humans in this work. We rely on existing datasets like CAESAR and AMASS together with the SMPL-X body model. These are all widely established datasets and models.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We use no human subjects or crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.