SYNPLERHUMAN: SIMPLIFIED SYNTHETIC DATA GENERATION FOR 3D HUMAN POSE AND SHAPE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic training data is widely used for 3D human pose and shape estimation and recent datasets invest large efforts in increasingly realistic humans, clothing, and environments. However, these improvements are usually introduced together, making it difficult to isolate which factors matter most for real-world generalization. To remedy this problem, we present SynplerHuman, a controllable, open-source procedural generator of data for 3D human pose and shape estimation". Our generator procedurally generates bodies, clothing, materials, and scene variations, enabling unlimited data generation without relying on external curated 3D appearance asset libraries. Using SynplerHuman, we study what factors of realism matter in synthetic data generation. We find that articulation choices have the largest impact on performance, while appearance variations produce relatively small changes. Finally, we demonstrate that SynplerHuman produces data that is effective when combined with existing datasets, and is competitive with synthetic datasets such as SynBody.

1 INTRODUCTION

Synthetic data is a practical way to obtain large scale supervision for 3D human pose and shape. It provides exact labels, controlled variation and the ability to generate training distributions that would be difficult to capture in the real world. As such, synthetic 3D human data has seen wide adoption in human pose and shape prediction methods. (Varol et al., 2017; Pumarola et al., 2019; Yang et al., 2023; Patel et al., 2021; Cai et al., 2024; Black et al., 2023; Bazavan et al., 2022; Yim et al., 2025; Yin et al., 2024; Tesch et al., 2025; Hewitt et al., 2024; Saleh et al., 2025; Ge et al., 2025; Jiang et al., 2024; Khirodkar et al., 2024). The realism of such data has progressed significantly. While early pipelines such as AGORA (Patel et al., 2021) had limited realism, modern systems leverage physically based rendering in diverse 3D environments, with rich asset libraries of human clothing and bodies.

However, realism improvements between datasets and data generators are not controlled, making it difficult to identify which factors are most useful. New releases frequently combine changes across pose sampling, scene composition, camera behavior and scale. This makes developing new synthetic datasets difficult. One might spend large effort on components that greatly increase implementation complexity, but contribute only little to performance. As such, we study the following question: *what aspects of realism matter for downstream pose and shape prediction*?

Existing generation pipelines make studying this question particularly difficult. Because existing generators typically use curated libraries of clothing, textures and scene assets (Black et al., 2023; Tesch et al., 2025), it is hard to produce matched variants where only one factor changes. To address this challenge, we introduce SynplerHuman, a generator of 3D human pose data with controllable realism. SynplerHuman builds off of Infinigen and is fully procedural, meaning that all assets are generated entirely from code.

In our study, we focus on three main realism factors commonly seen in synthetic human pipelines: appearance, articulation/pose, and physical realism. We study appearance by varying clothing, skin, and texture fidelity. To examine the importance of articulation sampling, we experiment with various learned and rule-based articulation sampling approaches. We also study physical plausibility

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Our procedural pipeline can generate variants across appearance, pose realism, scene context, or physical plausibility.

by randomizing global orientation and allowing humans to float in the rooms. In Figure 1 we show qualitative examples of variants that we consider in our study. We evaluate each variant by training CameraHMR Patel & Black (2025) and measuring performance across standard real-world benchmarks: 3DPW (von Marcard et al., 2018), RICH (Huang et al., 2022), EMDB (Kaufmann et al., 2023) and SPEC (Kocabas et al., 2021).

Our study finds that the most important parameter is articulation and pose sampling. Compared to a baseline of a learned human pose prior, we find that any changes to the pose sampling result in significant performance degradation. Although rule-based heuristics performed better than random sampling, all pose variations generally led to substantial performance degradation. Interestingly, we found that various appearance factors mattered relatively little, with unrealistic skin and naked humans having only a small negative effect on performance.

Finally, we demonstrate that our data is useful when compared to and in conjunction with existing synthetic datasets. In a data-matched comparison with SynBody (Yang et al., 2023), a mixed subset of our SynplerHuman-600k achieves a significantly lower error (56.93 vs 68.77 PA) on EMDB (Kaufmann et al., 2023). Moreover, our data also remains helpful when mixed with existing synthetic datasets. We find that replacing a subset of BEDLAM2.0 (Tesch et al., 2025) data can improve performance, even when using unrealistic data variants such as floating humans or random skin textures.

In summary our contributions are:

1. SynplerHuman, an open-source controllable procedural synthetic human pipeline with no dependence on external curated 3D human asset libraries.
2. An empirical analysis of which data factors are most influential for zero-shot sim2real transfer in 3D human pose and shape estimation.
3. SynplerHuman-600k, a dataset of 600,000 synthetic human images with ground truth generated in 4780 unique 3D scenes, covering varying levels of realism.

2 Related Work

Precise 3D pose and shape ground truth is hard to acquire at scale. Synthetic pipelines can instead render images together with exact supervision in parametric body models such as SMPL or SMPL-X

(Loper et al., 2015; Romero et al., 2017; Pavlakos et al., 2019). Prior work differs mainly in how it models context, ranging from 2D background compositing to full 3D scene synthesis.

2D background compositing. SURREAL renders SMPL bodies driven by MoCap and composites them over static indoor images sampled from LSUN (Varol et al., 2017). 3DPeople renders clothed humans with explicit body and clothing meshes and randomizes clothing textures, lighting and background images (Pumarola et al., 2019). Other works expand on the SURREAL idea but synthesize human sequences with posed garments with physics based deformation and varying backgrounds (Liu et al., 2019; Liang & Lin, 2019; Jiang et al., 2020). These pipelines treat scene context as an image layer. In our work, humans are rendered inside 3D scenes and we vary the context within the same synthetic generator.

3D scene composition and camera motion. AGORA renders multiple scanned people per image in natural scenes and provides SMPL-X ground truth together with occlusion aware segmentation masks(Patel et al., 2021). SPEC introduces SPEC-SYN, a photorealistic synthetic dataset with ground truth bodies and camera parameters designed to study perspective effects(Kocabas et al., 2021). SynBody constructs clothed parametric humans in a layered representation and renders them in Unreal Engine environments with layered annotations for body and clothing(Yang et al., 2023). HSPACE animates scan-based humans with GHUM and places them in complex indoor and outdoor environments with multi-person occlusion and camera motion(Bazavan et al., 2022). GTA-Human leverages the GTA-V engine to collect large scale video sequences with automatic SMPL annotations across diverse locations and actions(Cai et al., 2024). SynPlay targets the aerial view setting and uses rule guided motion generation (Yim et al., 2025). For world-grounded human and camera trajectories, WHAC-A-Mole and HCM provide synthetic data with ground truth camera motion(Yin et al., 2024; Kocabas et al., 2023). These systems often render into finite environment collections or rely on proprietary engines and purchased assets (Patel et al., 2021; Cai et al., 2024; Bazavan et al., 2022). By building on Infinigen, we obtain unbounded procedural scene variation and can regenerate matched variants by changing explicit parameters instead of sampling from a fixed scene catalog (Raistrick et al., 2023; 2024).

High fidelity and targeted synthesis. BEDLAM adds physics simulated clothing to SMPL-X humans, renders them in realistic scenes and reports strong real image performance when training on synthetic data alone (Black et al., 2023). BEDLAM2.0 expands camera motion diversity, improves hair assets and adds shoes while providing ground truth camera trajectories to support world-coordinate estimation (Tesch et al., 2025). Zolly and BLADE focus on close range imagery with strong perspective distortion and introduce synthetic datasets specific to this setting (Wang et al., 2023; 2024). VOccl3D targets realistic occlusions in video and provides ground truth pose and shape under heavy occlusion (Garg et al., 2025). Synthetic data also provides large benefits beyond pose regression (Saleh et al., 2025; Hewitt et al., 2024). These pipelines emphasize fidelity and often couple many design choices at once, such as camera behavior, clothing assets and scene libraries (Black et al., 2023; Tesch et al., 2025). Our main goal is different. We use a single procedural generator to create matched variants where one realism factor is modified at a time, which supports controlled comparisons.

Generative models. HumanWild uses diffusion models as data generators and conditions generation on keypoints, depth and surface normals rendered from SMPL-X (Ge et al., 2025). STAGE uses a text-to-image generator conditioned on SMPL renderings to produce paired benchmarks that change a single attribute while keeping the underlying 3D pose fixed (Kister et al., 2024). Both approaches reduce dependence on curated 3D assets, but they rely on a learned generator and additional steps to enforce label quality (Ge et al., 2025). Our work shares the goal of controlled comparisons, but uses renderer based synthesis so the ground truth is obtained directly.

Procedural Generation.Procedural generation replaces curated asset libraries with stochastic rules and explicit parameters. Infinigen generates photorealistic worlds fully procedurally and Infinigen Indoors extends it to indoor scenes with a constraint based arrangement system (Raistrick et al., 2023; 2024; Joshi et al., 2025). Our work follows the procedural approach for both scenes and humans and uses it to create matched dataset variants where one realism factor is modified while the rest of the pipeline is unchanged.



Figure 2: **SynplerHuman Dataset**. We random, non-cherrypicked examples from the SynplerHuman dataset. Refer to Appendix 8 for additional samples.

3 SynplerHuman Data Generator

In order to study dataset design, we first develop a procedural generation pipeline for human pose and shape data. We build our data generation pipeline on Infinigen (Raistrick et al., 2024), and implement an extended human generation stage and a configurable generation factory. Our system therefore enables (1) fully automatic large scale generation, (2) complete supervision for pose-and-shape learning and (3) controlled realism interventions.

Procedural scene and human generation Each scene is procedurally generated and then populated with SMPL-X humans in Blender. Human creation is implemented through a separate Infinigen stage, which controls body instantiation, appearance options, pose sampling, motion policy, clothing setup and annotation export. The number of humans per scene is sampled within a configured range (up to twelve in our experiments) and placement is validated to reduce collisions. Human bodies, skin and clothing appearance options and scene context are generated through procedural components, which lets us regenerate matched variants by changing the generator settings.

Motion and placement policies We use two motion strategies in the same pipeline. For grounded humans, start/goal locations are sampled from floor free space and connected with grid based A star paths. For floating humans, we use Infinigen's RRT based policy to find random trajectories in free space. In both cases, trajectory execution includes collision checks against static geometry and already placed humans and unsuccessful samples are resampled. For pose sampling we also utilize VolumetricSMPL (Mihajlovic et al., 2025) to generate pose sequences without self-collisions.

Appearance and Clothing Appearance is controlled procedurally through skin shading and clothing flags. For the SynplerHuman baseline, we reuse realistic human skin textures from (Black et al., 2023). Our skin ablation replaces these textures with intentionally non-realistic procedural materials. More specifically for these experiments we render 1) clothed humans, 2) unclothed humans and 3) variants with intentionally non-realistic skin or clothing materials.

Clothing is generated from the human mesh using vertex level segmentation, fully automatically without any need for post processing or manual alignment. Most of the clothing types are created by duplicating the mesh body parts segmentations, stitching them together and offsetting them from the human body. Additionally, we introduce a procedural skirt generator. The generated garments are optionally settled with a short cloth simulation and then bound to the body via Blender's surface deform modifier and pinning. To increase the diversity, we also create 5 different clothing presets

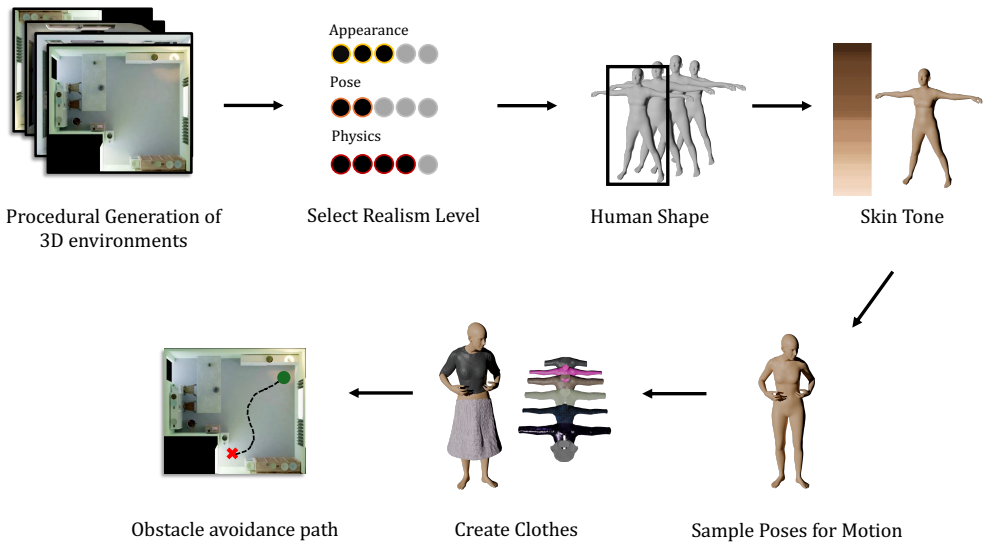


Figure 3: Overview of the procedural pipeline used in this work: scene generation, choosing the desired realism, human selection, motion/pose sampling, clothing generation, placing the humans in the scene and rendering and annotation export.

that are assigned to garments upon creation and affect the physics cloth simulation. See Figure 4 for example garment types and materials.

This clothing module is procedural and integrates directly with body instantiation and animation. In practice, this lets us vary clothing presence, outfit template and material appearance while keeping pose sampling, motion planning and rendering supervision fixed.

Annotation export For every rendered frame, the pipeline exports supervision needed by SMPL/SMPL-X based estimators: SMPL-X parameters, vertex and joint level 3D points, camera intrinsics/extrinsics and projected image space supervision (2D joints and 2D bounding boxes). For each human, we store a per-frame archive following similar structure as (Black et al., 2023; Tesch et al., 2025).

3.1 Pose Sampling and Articulation

Pose keyframes are sampled and interpolated through time to produce full animation sequences. In the SynplerHuman baseline body poses are sampled from DPoser (Lu et al., 2025), a diffusion based learned prior. Our sampler interface replaces this prior with multiple strategies, including unconstrained joint-range sampling, a per-joint Gaussian fit and rule-based samplers (normal and uniform). Global root orientation is sampled separately from local articulation to enable direct orientation prior ablations.

Per-joint Gaussian fit Given a reference pose bank (in our experiments we generate a pose bank using (Lu et al., 2025)), we fit a separate 3D Gaussian to each joint's rotation distribution (in a local axis-angle space around the mean rotation). At generation time we sample joints independently from these Gaussians and we inflate the covariance to increase diversity. This matches per-joint marginals but ignores inter-joint correlations.

Rule-based samplers We implement a kinematic prior with explicit anatomical joint ranges. Hinge joints (e.g., knees and elbows) are parameterized by flexion plus limited twist and lateral deviation. Ball joints (e.g., hips and shoulders) are parameterized by a swing direction inside a cone and a twist about the bone axis. The *normal* variant samples these parameters from truncated Gaussians within the allowed range, while the *uniform* variant samples uniformly within the same limits. The sampler also supports optional symmetry coupling for hips and shoulders and a bias toward arms down shoulder directions for our heuristic variants.



(a) Examples of procedural clothing outfits fit to variety of different body shapes without manual alignment.

(b) Examples of procedural clothing materials. Each garment is assigned a random procedural material and sampled from one of 5 predefined clothing presets.

Figure 4: Procedural clothing variation used in the pipeline, covering both outfit type combinations and material appearance diversity.

3.2 Controlled Realism Axes

Using the same base generator, we create matched variants that modify one factor at a time:

1. Appearance: remove clothing, degrade skin textures, degrade clothing realism.
2. Articulation: unconstrained joint ranges, per-joint Gaussian fitted statistics, rule based normal/uniform samplers and rule based approach with heuristics.
3. Physical plausibility: floating humans, random global orientation

We generate samples across each variant and combine them to create SynplerHuman-600k, a dataset of 600,000 SMPL-X human instances across 4,780 unique procedurally generated Infinigen scenes, organized into the controlled realism variants in this paper. Each scene is rendered as a short clip (frames 0-40). For every frame we provide full supervision in SMPL-X space (pose, shape and global orientation), camera intrinsics/extrinsics, 3D joints and mesh vertices and projected 2D keypoints and bounding boxes. To get a variety of different human shapes, we uniformly sample the shape parameter β for every instantiated human. For clothed variants, outfits are sampled from 40 procedural templates built from our garment catalog (tops, skirts, pants, shorts etc.). Each garment is further assigned a cloth material preset (we created 5 different clothing presets) that controls cloth settling parameters and a procedural surface material for appearance diversity. We extended the existing Infinigen fabric materials with three additional procedural materials to increase the visual variation.

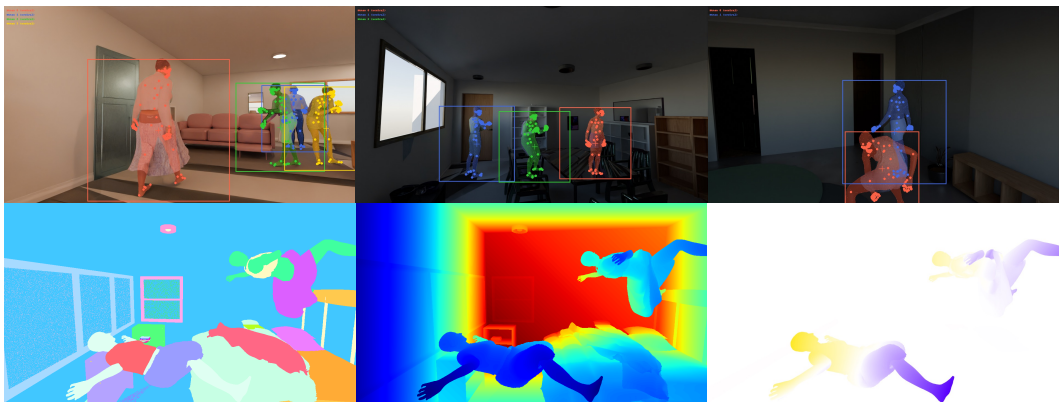


Figure 5: Example ground truth the pipeline can produce. We utilize Infinigen's export to obtain ground truth like depth, segmentation or optical flow and additionally export parametric human ground truth, camera intrinsics/extrinsics and 2D projections.

	EMDB			3DPW			SPEC			RICH			
Variant	PA	MPJPE	PVE	PA	MPJPE	PVE	PA	MPJPE	PVE	PA	MPJPE	PVE	Avg PA
Procedural baseline	**63.2**	96.8	111.1	49.9	84.8	97.6	**56.7**	86.0	98.8	**57.3**	**89.5**	**104.0**	**56.8**
Floating humans	64.0	**96.0**	**110.1**	**48.8**	**75.1**	90.0	57.1	86.8	101.4	60.2	94.0	107.0	57.5
Non-realistic skin textures	64.8	101.6	116.3	51.7	88.2	99.7	58.0	88.3	102.0	58.7	92.4	108.9	58.3
No clothing	66.8	104.2	122.4	50.6	75.7	**89.7**	58.5	**85.7**	**98.5**	60.3	93.6	110.3	59.1
Non-realistic clothing	67.3	102.4	119.3	54.4	80.5	96.3	64.6	93.7	110.1	65.2	95.2	113.0	62.9
Random pose orientation	70.7	105.8	121.0	59.8	101.9	113.7	65.4	105.6	122.6	64.1	100.1	119.0	65.0
Per-joint Gaussian fit	76.0	113.7	132.7	69.6	96.3	114.5	87.9	113.1	133.3	70.1	103.6	124.0	75.9
Rule-based pose (normal)	81.6	117.5	135.3	75.2	103.5	121.5	98.0	124.0	145.8	70.6	107.5	126.3	81.3
Rule-based pose (uniform)	92.2	127.1	147.9	86.8	113.3	134.3	113.9	135.2	154.9	80.2	117.6	136.1	93.3
Random joint pose	173.7	230.7	280.4	171.2	218.9	243.2	187.1	223.5	255.7	150.8	199.4	241.0	170.7

Table 1: **Ablation Study.** All metrics in mm, lower is better. PA denotes PA-MPJPE; Avg PA is the mean PA across the four datasets. Rows are sorted by ascending Avg PA. Best per column in **bold**.

4 Experiments

Training Each experiment is trained on one synthetic variant at a time and evaluated with the same protocol. We keep architecture, optimization and evaluation settings fixed across variants to isolate the effect of generator choices. For the core experiments we use a fixed per-variant training set size to keep the training budget constant. We train CameraHMR (Patel & Black, 2025) using the official codebase and adapt its regression target and supervision to predict SMPL-X parameters (Pavlakos et al., 2019). The model is trained with fully supervised synthetic ground truth (SMPL-X parameters, 3D joints, mesh vertices and 2D keypoints). We follow the default CameraHMR preprocessing and augmentation pipeline and keep loss weights and optimizer settings fixed across all variants.

Benchmarks and Metrics We evaluate performance on four popular image benchmarks: 3DPW (von Marcard et al., 2018), EMDB (Kaufmann et al., 2023), RICH (Huang et al., 2022) and SPEC (Kocabas et al., 2021). These datasets provide complementary test conditions, including indoor and outdoor scenes, static and moving cameras, and varying levels of occlusion. To measure the differences between our different realism variants, we use commonly used metrics for human pose and shape estimation, namely: MPJPE, PA-MPJPE and PVE (all in millimeters and lower is better).

5 Results

5.1 Parameter Study

Table 1 summarizes the results from the different realism variants. Each dataset column is reported as PA-MPJPE/MPJPE/PVE (mm, lower is better) and the rightmost column summarizes average PA-MPJPE across the four benchmarks. In Figure 6 we show qualitative results on different difficult real world examples for a subset of the variants.

Articulation. The largest changes in performance come from how we sample articulation. Replacing the learned prior in our baseline with unconstrained joint-range sampling increases average PA-MPJPE from 56.77 to 170.71. Handcrafted priors and rules reduce this gap but still remain worse than utilizing the learned human pose prior, including per-joint Gaussians (75.90), rule-based normal (81.33) and rule-based uniform (93.27). In our sweep, articulation sampling is the dominant source of variation in sim-to-real transfer.

Appearance. Appearance only changes lead to smaller differences. If we degrade skin textures with procedural materials, we observe that it increases the average PA-MPJPE by +1.54 compared to the baseline. Similarly, removing clothing increases error by +2.28. In contrast, using non-realistic looking clothing items (E) is more harmful and leads to an increase of +6.09.

Setting	EMDB	3DPW	SPEC	RICH	Avg PA
Rule based uniform	92.19	86.79	113.94	80.17	93.27
Rule based uniform	78.72	74.08	90.05	66.20	77.26
Rule based uniform (XL model)	75.81	69.59	87.65	65.97	74.75
Rule based uniform + heuristics	81.51	70.03	85.66	72.77	77.49
Rule based normal + heuristics	82.43	72.13	88.60	73.94	79.28
Procedural baseline (DPoser)	63.2	49.9	56.7	57.3	56.8

Table 2: **Scaling rule-based articulation**. Scale and heuristic runs reported as PA-MPJPE (mm, lower is better). Scaling up rule-based articulation cannot overcome a learned pose prior.

Grounding and orientation. Violating physical grounding by allowing floating humans (H) changes performance only slightly (57.51) vs. (56.77). Similarly to articulation, randomizing global orientation (L) is more detrimental to the HPS performance (65.00).

5.2 Scaling Heuristic Pose Sampling

We would like to study whether scale and simple heuristics could compensate for weaker pose priors. As such, we evaluate two larger scale variants of the rule based uniform sampler using 260K SMPL-X human instances, one on a standard CameraHMR model architecture (Patel & Black, 2025) and one on adjusted architecture where the size of the head is increased to roughly 1.45B (denoted as XL model in Table 2). Additionally, we experiment with two heuristic biased variants that encourage more common arms down postures and mirrored limbs. We denote scaled runs by `+S` and heuristic biased runs by `+H` in Table 2. We find that scaling the rule-based uniform variant improves average PA-MPJPE from 93.27 to 77.26 and increasing model capacity gives a further improvement to 74.75. Adding symmetry and arms down heuristics gives a similar gain (77.49). Nevertheless, our baseline with a learned pose prior still performs better than even the scaled up runs.

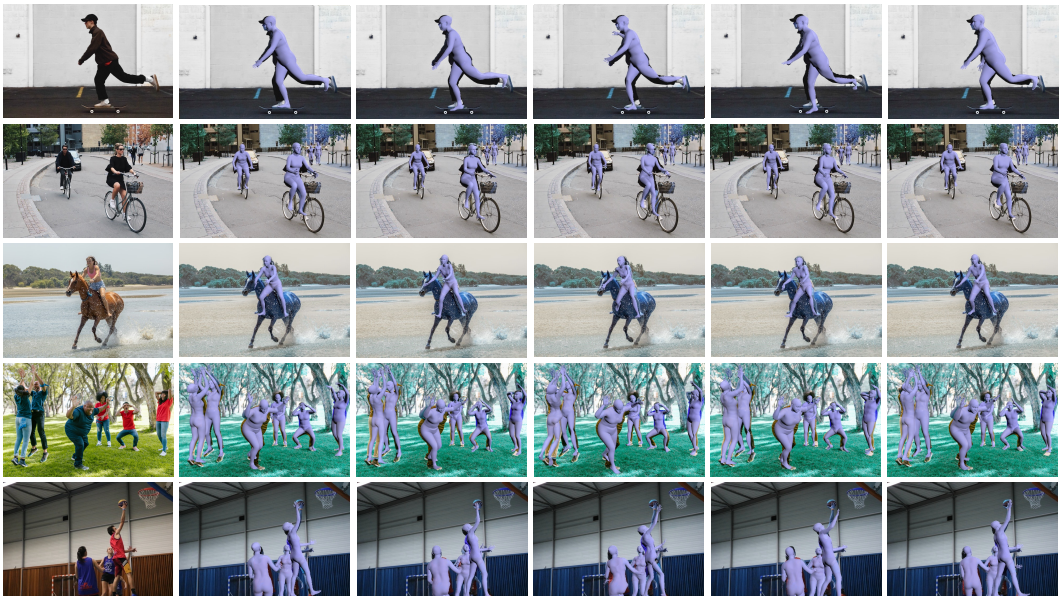


Figure 6: Qualitative examples of models trained on different realism variations shown on challenging human activities.

Training set	Crops	PA	MPJPE	PVE
BEDLAM2	101,557	**50.83**	**85.03**	**98.01**
SynplerHuman B+C+D+E+H	101,468	56.93	91.49	105.54
SynBody	101,468	68.77	109.91	128.30
AGORA	86,778	66.39	135.02	154.51

Table 3: **Comparison with existing datasets.** We train CameraHMR on varying existing synthetic datasets using a fixed data budget and evaluate on EMDB. Our SynplerHuman data outperforms SynBody and AGORA, while being fully procedural.

+11k crops of	PA	MPJPE	PVE	Δ vs. control
SynplerHuman baseline	**53.38**	**85.09**	**100.02**	$\mathbf{-1.61}$
non-realistic clothes	53.93	85.99	101.65	-1.06
procedural skins	54.31	86.29	101.09	-0.68
floating humans	54.90	90.71	106.73	-0.09
control: +11k more BEDLAM2	*54.99*	*90.21*	*106.52*	*0.00*
naked humans	55.02	88.46	104.46	$+0.03$
rule-based uniform poses	55.79	87.89	104.79	$+0.80$
per-joint Gaussian poses	55.87	89.85	106.31	$+0.88$
rule-based heuristic	56.01	88.89	105.67	$+1.02$
random poses	56.62	91.18	108.21	$+1.63$
random global orientation	56.81	89.49	105.78	$+1.82$
base: BEDLAM2 (35,911 crops)	*56.63*	*90.76*	*106.31*	*+1.64*

Table 4: **Data mixing experiments.** Given an initial dataset of 36k BEDLAM2.0 crops, we ablate which dataset types are most effective when added to the data mix. We find that many unrealistic variants, such as procedural skins, can actually perform better than the control of simply adding more BEDLAM2.0 data. Metrics are on EMDB.

5.3 Dataset Evaluation and Mixing

Comparison with existing datasets. We compare the performance of various datasets alone in a data-matched setting. We train CameraHMR on a fixed budget of roughly 101k crops from BEDLAM2.0, SynBody and SynplerHuman, and on 86.8k crops from AGORA. We evaluate on EMDB and report the results in Table 3. We construct the SynplerHuman mix by drawing from five variants that use the learned pose prior: the baseline, no clothing, non-realistic skin, non-realistic clothing and floating humans. We find that this mix is able to achieve significantly lower error than SynBody (56.93 vs 68.77 PA), but is not able to surpass BEDLAM2.0 in overall performance.

Mixing with BEDLAM2.0. Starting from 35.9k BEDLAM2.0 crops, we add 11k crops from a single SynplerHuman variant and compare against a control of adding 11k more BEDLAM2.0 crops (54.99), as shown in Table 4. We find that non-realistic data can actually be more effective in the mixed data regime, with settings such as floating humans and procedural skins performing better than realistic BEDLAM data. However, as seen in the original parameter study, the best performing data is still the realistic SynplerHuman baseline, while the worst performing data is data with degraded articulation.

6 Conclusion

We presented a controlled study of realism factors for synthetic training data in 3D human pose and shape estimation. Using a procedural pipeline, we generated matched realism variants and evaluated transfer on 3DPW, RICH, EMDB and SPEC. Articulation and orientation drove the largest performance differences, while several appearance degradations were comparatively less impactful. Finally, we demonstrate that SynplerHuman data is effective both independently for training, and when mixed with other existing datasets. We hope that these results provide practical guidance for efficient synthetic human dataset design.

AI USE STATEMENT

We used LLMs to to aid or polish writing; for instance, we checked for grammatical errors in the manuscript with LLMs. We used LLMs for research execution, and wrote some portions of the experimental code using LLMs.

REPRODUCIBILITY STATEMENT

We will open source our dataset generator, training code, and all data used in the ablation experiments.

REFERENCES

Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Hspace: Synthetic parametric humans animated in complex environments, 2022. URL https://arxiv.org/abs/2112.12867.

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

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*, pp. 1–12, 2024. doi: 10.1109/TPAMI.2024.3450537.

Yash Garg, Saketh Bachu, Arindam Dutta, Rohit Lal, Sarosij Bose, Calvin-Khang Ta, M. Salman Asif, and Amit Roy-Chowdhury. Voccl3d: A video benchmark dataset for 3d human pose and shape estimation under real occlusions, 2025. URL https://arxiv.org/abs/2508.06757.

Yongtao Ge, Wenjia Wang, Yongfan Chen, Fanzhou Wang, Lei Yang, Hao Chen, and Chunhua Shen. Diffusion models are efficient data generators for human mesh recovery, 2025. URL https://arxiv.org/abs/2403.11111.

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.

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 *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 13274–13285, June 2022.

Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image, 2020. URL https://arxiv.org/abs/2004.00214.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1737–1747, 2024.

Abhishek Joshi, Beining Han, Jack Nugent, Max Gonzalez Saez-Diez, Yiming Zuo, Jonathan Liu, Hongyu Wen, Stamatis Alexandropoulos, Karhan Kayan, Anna Calveri, Tao Sun, Gaowen Liu, Yi Shao, Alexander Raistrick, and Jia Deng. Procedural generation of articulated simulation-ready assets, 2025. URL https://arxiv.org/abs/2505.10755.

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.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024.

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. URL https://arxiv.org/abs/2408.16536.

Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proceedings International Conference on Computer Vision (ICCV)*, pp. 11035–11045. IEEE, October 2021.

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, 2023. URL https://arxiv.org/abs/2310.13768.

Junbang Liang and Ming C. Lin. Shape-aware human pose and shape reconstruction using multi-view images, 2019. URL https://arxiv.org/abs/1908.09464.

Jian Liu, Naveed Akhtar, and Ajmal Mian. Temporally coherent full 3d mesh human pose recovery from monocular video, 2019. URL https://arxiv.org/abs/1906.00161.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), November 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818013. URL https://doi.org/10.1145/2816795.2818013.

Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, and Ziwei Liu. Dposer-x: Diffusion model as robust 3d whole-body human pose prior. *arXiv preprint arXiv:2508.00599*, 2025.

Marko Mihajlovic, Siwei Zhang, Gen Li, Kaifeng Zhao, Lea Müller, and Siyu Tang. VolumetricSMPL: A neural volumetric body model for efficient interactions, contacts, and collisions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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

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 *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

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, 2019. URL https://arxiv.org/abs/1904.05866.

Albert Pumarola, Jordi Sanchez, Gary P. T. Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans, 2019. URL https://arxiv.org/abs/1904.04571.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12630–12641, 2023.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21783–21794, June 2024.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

Fatemeh Saleh, Sadegh Aliakbarian, Charlie Hewitt, Lohit Petikam, Xiao-Xian, Antonio Criminisi, Thomas J. Cashman, and Tadas Baltrušaitis. DAViD: Data-efficient and accurate vision models from synthetic data, 2025. URL https://arxiv.org/abs/2507.15365.

Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J. Black. BEDLAM2.0: Synthetic humans and cameras in motion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017.

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 *European Conference on Computer Vision (ECCV)*, sep 2018.

Shengze Wang, Jiefeng Li, Tianye Li, Ye Yuan, Henry Fuchs, Koki Nagano, Shalini De Mello, and Michael Stengel. Blade: Single-view body mesh learning through accurate depth estimation, 2024. URL https://arxiv.org/abs/2412.08640.

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, 2023. URL https://arxiv.org/abs/2303.13796.

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)*, pp. 20282–20292, October 2023.

Jinsub Yim, Hyungtae Lee, Sungmin Eum, Yi-Ting Shen, Yan Zhang, Heesung Kwon, and Shuvra S. Bhattacharyya. Synplay: Large-scale synthetic human data with real-world diversity for aerial-view perception, 2025. URL https://arxiv.org/abs/2408.11814.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, Ziwei Liu, and Lei Yang. Whac: World-grounded humans and cameras, 2024. URL https://arxiv.org/abs/2403.12959.

Method	Learned	Real data	Anatomy	Corr.
Unconstrained sampling	✗	✗	✗	✗
Gaussian (independent)	✗	✓	✗	✗
Gaussian (correlated)	✗	✓	✗	✓
Rule-based (uniform)	✗	✗	✓	✗
Rule-based (normal)	✗	✗	✓	✗
Learned Prior	✓	✓	✓	✓

Table 5: Comparison of pose sampling strategies that can be generated by the synthetic pipeline. **Anatomy** indicates explicit kinematic constraints and **Corr.** denotes modeling of inter-joint correlations. **Real data** indicates that the pose sampling uses real pose data to fit the distribution.

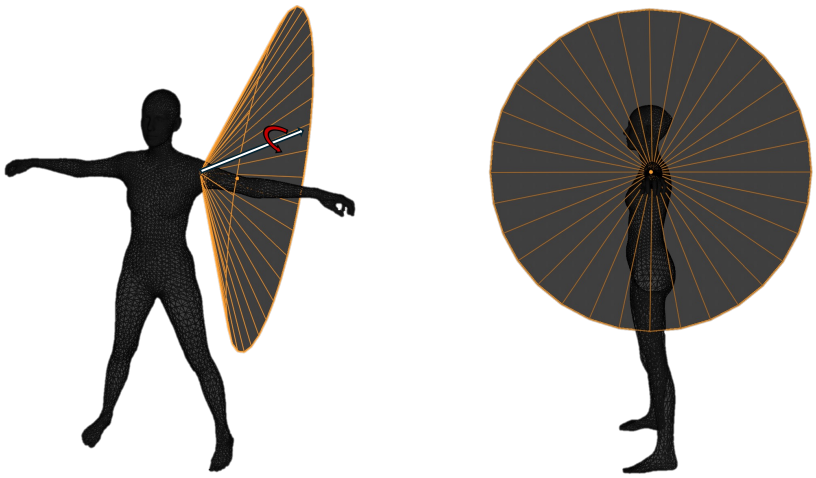


Figure 7: Rule-based pose sampling used for Variants I and J. Ball joints are parameterized by swing inside a joint-specific cone plus twist around the rotated bone axis.

A Appendix

A.1 Additional Implementation Details

A.1.1 Training Configuration

All core experiments in the paper train the same CameraHMR model on one SynplerHuman dataset variant. We adapt the official CameraHMR training code to regress SMPL-X parameters and supervise the model with synthetic SMPL-X pose and shape parameters, 3D joints and 2D keypoints.

CameraHMR is optimized with AdamW using learning rate 10^{-5}, weight decay 10^{-4} and batch size 32. The ViTPose backbone is initialized the same way as the original work.

A.1.2 Pose Generation and Filtering

For articulated pose realism we study four families of samplers. The baseline uses a learned pose prior. The Gaussian variant fits per-joint Gaussians in local rotation space to a reference poses. The rule-based samplers explicitly encode joint limits for hinge and ball joints and differ only in whether latent parameters are sampled from truncated normal or uniform distributions. The weakest baseline ignores joint correlations and samples independent joint rotations within broad anatomical ranges. Table 5 summarizes the main assumptions behind these pose sampling strategies. See an example of how the rule-based sampler works in Figure 7.

To reduce obviously invalid samples, candidate poses are filtered with a self-intersection check. We reject poses whose collision score exceeds a fixed threshold and resample until a valid pose is obtained.

Variant	Axis	Modified component
B	Baseline	Procedural pipeline with realistic clothing, learned pose prior, grounded placement and realistic root-orientation prior
C	Appearance	Clothing removed
D	Appearance	Realistic skin replaced by intentionally non-realistic procedural skin materials
E	Appearance	Clothing replaced by intentionally non-realistic appearance
F	Articulation	Learned pose prior replaced by weakly constrained random joint-range sampling
G	Articulation	Learned pose prior replaced by a per-joint Gaussian fitted to a reference pose bank
H	Physical plausibility	Humans may float instead of being constrained by grounding/contact policies
I	Articulation	Learned pose prior replaced by rule-based sampling with truncated-normal latent sampling
J	Articulation	Learned pose prior replaced by rule-based sampling with uniform latent sampling
L	Orientation prior	Global root orientation randomized independently of the realistic upright prior

Table 6: Detailed definition of the realism variants used in the controlled study. For every non-baseline variant, all unspecified components remain identical to Variant B, including annotation export, CameraHMR training setup, shape sampling, data budget and evaluation protocol.

A.2 Variant Definitions

See Table 6 for a detailed definition of the realism variants used in the controlled study. For every non-baseline variant, all unspecified components remain identical to Variant B.

A.3 Additional Qualitative Examples

Figure 8 shows more examples from SynplerHuman and the realism variants. Figure 9 shows human heads rendered with different skin materials used in our study. Figure 10 illustrates how we apply cloth simulation and physics to procedural garments. Figure 11 shows our segmentation of human body parts that we used for garment construction.

A.4 Compute Resources

Data generation was performed on NVIDIA RTX 3090 GPUs. Training and evaluation used an LSF-managed GPU cluster with NVIDIA V100 32GB GPUs. Each reported training or evaluation job requested one V100 GPU in exclusive-process mode, 4 CPU cores, 32GB system memory, a single host, and a 24 hour wall-clock limit. We report single-run results for each variant rather than multi-seed averages. The reported sweep contains 10 core realism variants and 5 scale or heuristic runs; using the job wall-clock limit gives an upper-bound training/evaluation budget of approximately 360 V100 GPU-hours for the reported runs, plus the RTX 3090 GPU time used for synthetic data generation.

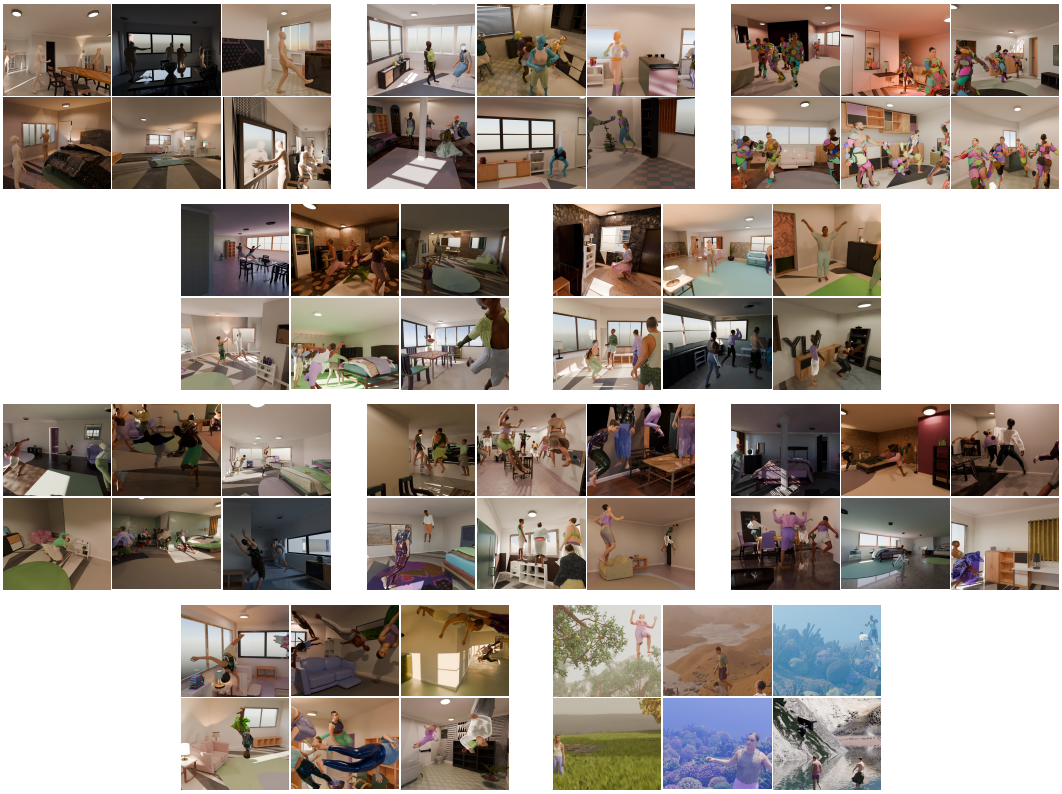


Figure 8: More examples from SynplerHuman variants.



Figure 9: Human heads rendered with different procedural skin materials.



Figure 10: Example of applying cloth simulation and physics to shorts obtained from the human model thighs.



Figure 11: Our human segmentation of body parts used for garment construction.