Are Pose Estimators Ready for the Open World? STAGE: A GenAI Toolkit for Auditing 3D Human Pose Estimators

Supplementary Material

A. Implementation details

A.1. Training CN-3DPose

The CN-3DPose architecture is based on ControlNet [13], where we adapt the input layer to take 9 channels, 3 for each of the 3 input conditions (dense SMPL semantic encoding, 2D skeleton drawing, depth). We initialize the weights from a pre-trained CN-Depth [13] and train on AGORA [7], HUMBI [12], SHHQ [3], COCO [5] and a set of human scan renderings [1, 9, 11].

We use image crops of size 512×512 px. We remove images where the person is significantly occluded or truncated. For COCO, we require at least 10 visible keypoints; for AGORA we require 20% of SMPL vertices to project within the image boundary. When the person is near the edge of the image, we use zero padding to ensure the person is centered in the crop. We mask these padding regions in the loss computation, so the model does not learn to generate them at test time. Since we focus on single-person scenarios in this work, we also mask out other people's regions from the loss, using the masks provided in the datasets.

Caption generation. CN-3DPose is a pose-conditioned text-to-image model. To train it, we need image captions in the training data. We generate these using BLIP2 [4]. For each image, we sample five captions and filter them using CLIP [8] to ensure image-text alignment. For COCO, we also use the captions provided in the dataset annotations. During training, we randomly pick a caption from this filtered set for each image.

We use some renderings in our training data (AGORA and the scan renders), but would like to avoid reproducing their distinct CG look at test time. To this end, we add the word *"Rendering"* to the captions of rendering-based training images, and use negative prompting with this word at inference time.

Training details. For all hyperparameters, we follow [13]. For batch construction we sample examples from each dataset with probability proportional to the size of the dataset. We train on 4 NVIDIA A100 (40GB) for nine days.

A.2. 2D pose estimation-based filtering

To identify low-quality generations we use OpenPose [2] to predict 2D keypoints and compare them to the ground truth. We convert the SMPL-based ground-truth pose to the OpenPose/COCO format using a joint regressor on the SMPL vertices. The image is discarded if the projection error exceeds 50 px for at least one wrist, ankle, shoulder, elbow, or knee keypoint, or OpenPose detects multiple people in the image.

A.3. Replica of 3DPW

To validate our image quality, we create a synthetic replica of a subset of 3DPW. We sample 1500 examples based on farthest point sampling of the poses to obtain a diverse subset. We use the same camera angles as in 3DPW. To create the text prompts, we use the prompt template *"Photo, caucasian {gender} wearing {clothing} in {location} during a sunny day at daytime"*. We fill in the gender from the annotations, and use BLIP2 [4] to describe clothing and location, by asking *"What is the person in the foreground wearing?"* and *"Where is the person located?"*.

We use 2D pose-based and VQA-based quality checks, and regenerate low-quality samples. However, if we exceed 13 attempts for a given pose, we discard that pose. The number of resulting valid poses is 1021 for CN-Pose, 1367 for CN-Multi and 1372 for CN-3DPose. For fair comparison between them, we only use the common subset of these, containing 981 examples.

A.4. Data generation for attribute experiments

Camera settings. For the attribute robustness experiments, we use 1500 AMASS poses as described in the main paper. Since AMASS does not contain camera information, we construct the camera parameters as follows. The camera faces the root joint frontally, such that the wrists, ankles, shoulders, and elbows are visible and the distance is such that the pose fits tightly within the frame.

Base prompt. Generally we use the base prompt *"Photo, caucasian {gender} wearing t-shirt and pants in city center at daytime sunny weather"*, where gender is filled in based on the pose that is used for generation. However, for some specific experiments, we need to make adjustments. For the indoor location experiment, the base prompt has *"hallway"* instead of *"city center"* and in the outdoor location experiment we use *"village"* in the base prompt to evaluate *"city center"* itself. For gender, we use *"male"* in the base set and *"female"* for the attribute set. For body shape, we use *"adult with average BMI"* for base and *"adult with low/high BMI"* for the two attribute datasets.

Quality filtering. We again apply 2D pose and VQA-based quality checks and filtering. For fair comparison between attributes of a category, we use only poses that resulted in valid images for all attributes of that category.

	Location (outdoor)	Location (indoor)	Fairness	Clothing	Weather	Texture
# poses	1062	1320	660	708	370	1085

Table S1. **Number of poses used in the robustness experiments**. We use several hundred to more than a thousand examples to compute PDP for each attribute. The numbers differ between categories because we filter out low-quality generations. In each category, we only keep those poses for which generation was successful for every attribute of that category. This ensures that PDP can be compared reliably among the different attributes of a category.

The number of resulting poses is listed in Tab. S1. To make sure that this number of poses is enough for a reliable calculation of the percentage of degraded poses (PDP), we investigate how the PDP changes with different numbers of samples. As seen in Fig. S1, the PDP is stable after the first couple of hundred poses, confirming that the number of samples is sufficient.

B. BEDLAM vs. STAGE

Training data generation with CN-3DPose. Since BEDLAM is a video dataset, the poses for consecutive frames can be very similar. To reduce redundancy in the data, we therefore pick poses only if at least one joint has moved by at least 10 cm. We further discard highly occluded instances that cannot be detected by YOLOv4. This filtering results in a set of 355k annotated image crops. For each of these examples, we then generate a corresponding image using our CN-3DPose model using the pose annotation for conditioning. For the text prompt, we populate our prompt template from Sec. A.3 with random combinations of the attributes used in our robustness experiments. Similar to the robustness experiments, here we also perform quality checks based on OpenPose's 2D estimation, and discard generations where OpenPose has high error. The result is a dataset with the same body pose and camera distribution as BEDLAM, but with more diverse and more realistic images.

Pose estimator training details. We use MMHuman3D [6] to train HMR and PARE—both of them with the HRNet-W48 [10] backbone pretrained on COCO keypoints. We train for 50 epochs with a batch size of 128 for HMR and 64 for PARE. The learning rate starts at 3.3×10^{-5} for HMR and at 1.6×10^{-5} for PARE, then decays exponentially with a rate of 0.9729 for the first 40 epochs and 0.896 for the last 10 epochs. We use a single NVIDIA A100 (40GB) GPU.

C. Full results

The full results of our attribute experiments using STAGE are shown in Figs. S2 to S7. We also provide more visual examples in Fig. S8, showing the utility of STAGE to test pose estimators with many different attributes.

We observe that none of the pose estimators are completely robust against attribute changes and can change their prediction if one aspect of the image is changed. This is best observed in Figs. S2 and S4, where we see a continual increase of PDP across all estimators from left to right. This indicates that the more difficult attributes affect even the most robust models.

In the outdoor location experiments each location shows a similar impact, suggesting that they are equally challenging. Only *"swamp"* and *"wetlands"* show a slightly higher impact compared to the rest of the attributes. In contrast, the indoor locations differ from each other to a larger extent. The most difficult indoor locations (*"bar"*, *"restaurant"* or *"kitchen"*) are those where occlusions by tables are common. Indeed, we observe that many such images depict occluded limbs, which suggests the leading cause of error for these attributes to be occlusions.

D. Qualitative results

We compare our method CN-3DPose with CN-Pose, CN-Depth, and CN-Multi in terms of diversity in Fig. S9 by generating images with fixed input pose (per figure) and different prompts. CN-Depth and CN-Multi achieve good pose alignment but are not able to follow the prompt and generate only flat backgrounds. CN-Pose creates diverse images but does not provide good pose alignment. Overall, only our method, CN-3DPose, can generate diverse images while having good pose alignment.

References

[1] AXZY Dataset. Axyz dataset. https://secure.axyz-design.com. Accessed: 2024-03-07. S1

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE TPAMI*, 2019. S1

[3] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A data-centric odyssey of human generation. In *ECCV*, 2022. S1

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. S1

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. S1

[6] MMHuman3D. OpenMMLab 3D human parametric model toolbox and benchmark. https://github.com/open-mmlab/mmhuman3d, 2021. S2

[7] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black.

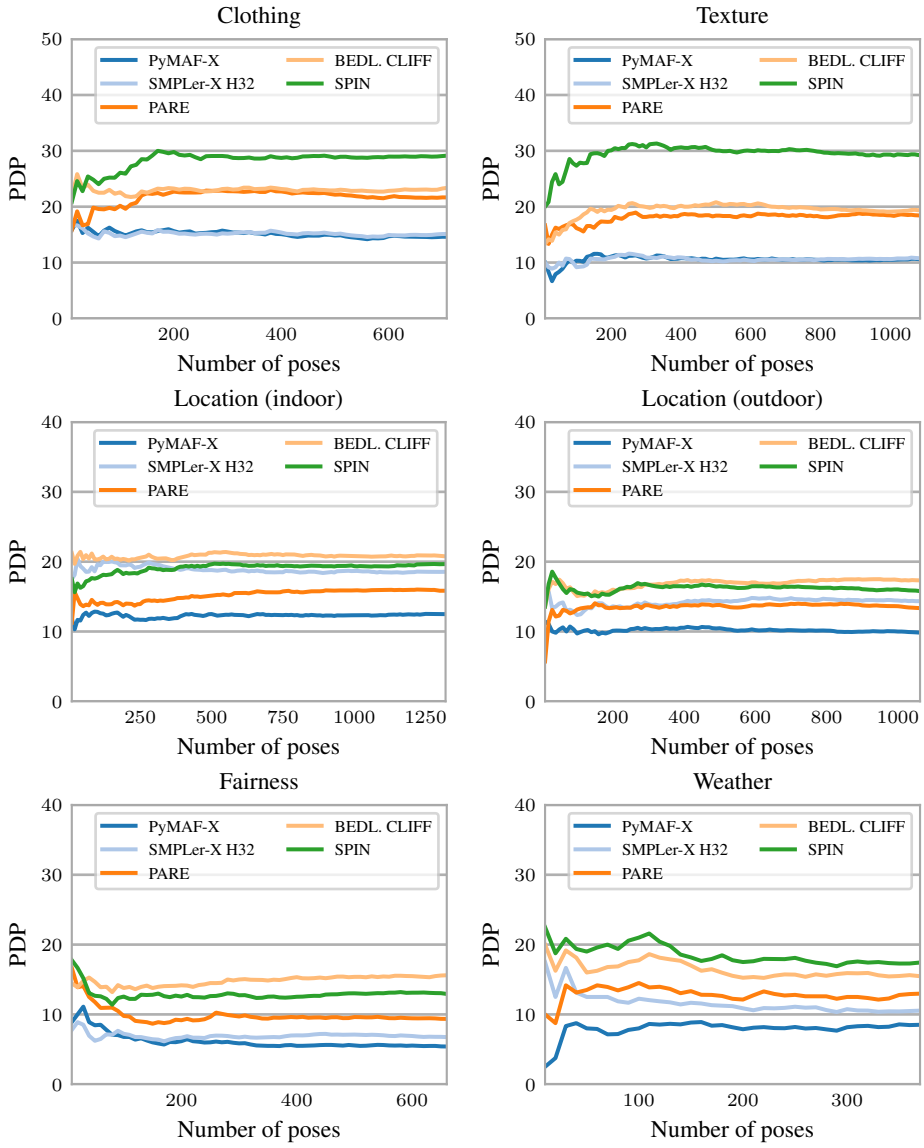

Figure S1. **Our number of samples is sufficient.** We plot how our results per category change based on the number of valid poses used to compute them. Overall the PDP remains stable after the first couple of hundred of poses.

AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, 2021. S1

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. S1

[9] RenderPeople Dataset. RenderPeople dataset. https://renderpeople.com, 2020. Accessed: 2024-03-07. S1

[10] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. S2

[11] Twindom Dataset. Twindom dataset. https://web.twindom.com. Accessed: 2024-03-07. S1

[12] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions. In *CVPR*, 2020. S1

[13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. S1

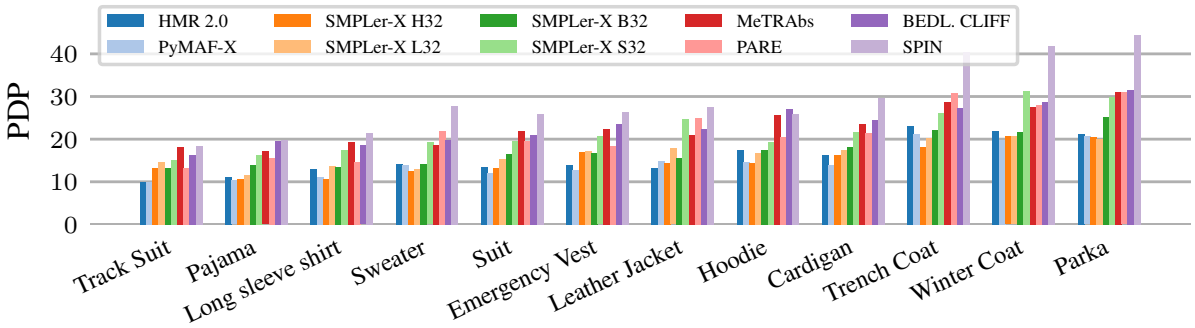

Figure S2. **Clothing impacts performance.** Clothing has the largest impact on performance. Especially items that cover most of the body, such as coats, can impact the performance in a negative way.

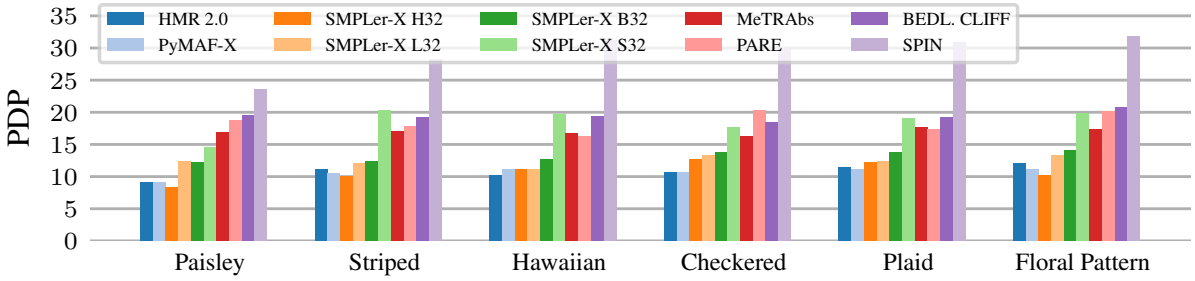

Figure S3. **Pose estimators are susceptible to texture changes.** All textures lead to about the same PDP, indicating that a texture change influences performance regardless of what that texture is. SPIN is particularly sensitive as up to 30 percent of the poses are degraded.

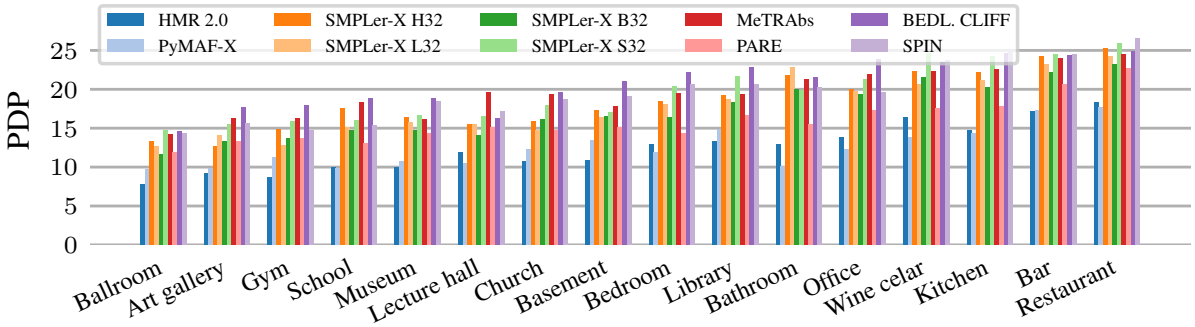

Figure S4. **Influence of indoor locations.** The continual increase of the PDP indicates that some locations are more challenging than others. Especially, “Restaurant”, “Bar” and “Wine cellar” have the most impact on performance.

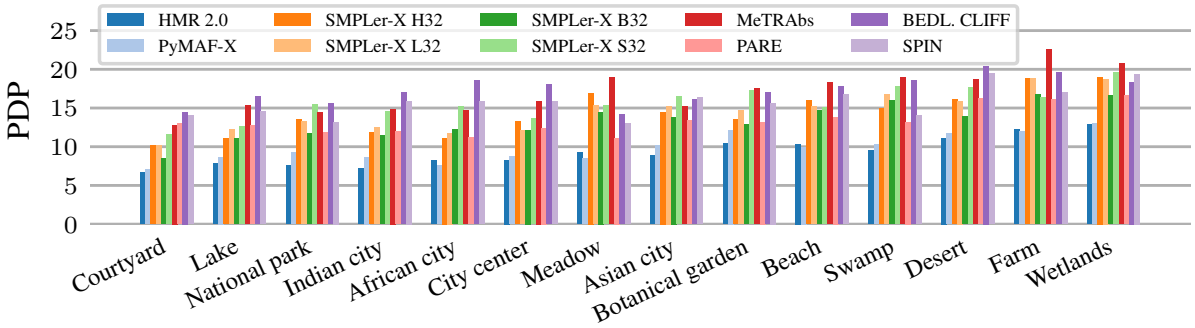

Figure S5. **Influence of outdoor locations**. Most outdoor locations have a similar impact on performance, “Swamp and “Wetlands” pose the most challenges to pose estimators.

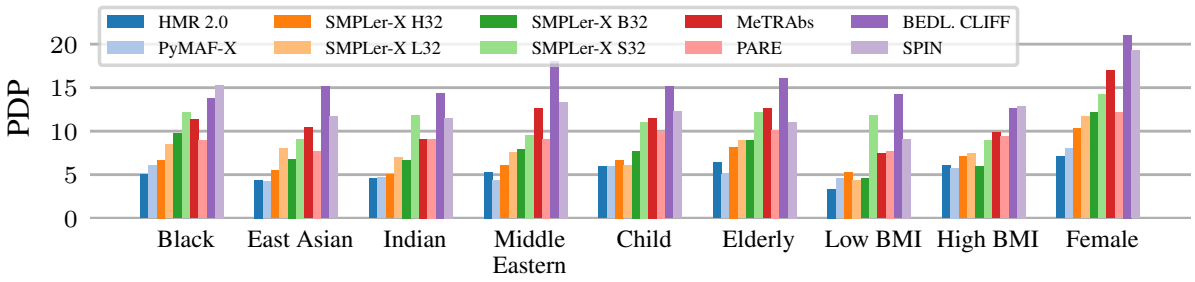

Figure S6. **Fairness analysis.** We consider multiple attributes related to fairness in computer vision. Pose estimators are robust against protected attributes. However, gender and age.

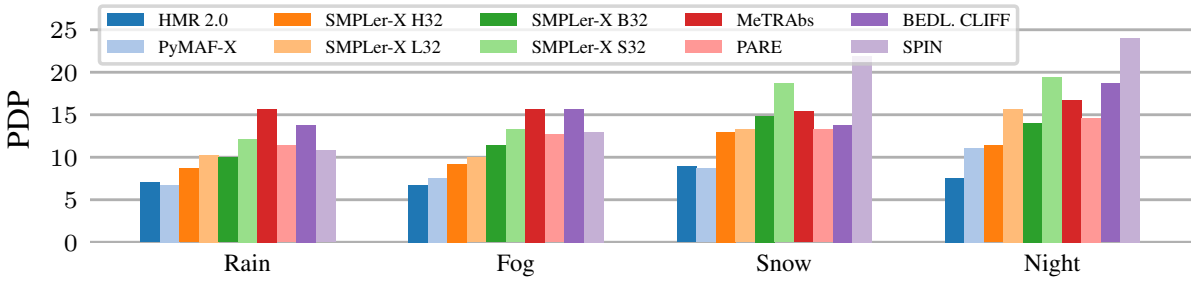

Figure S7. **Influence of adverse conditions**. Adverse conditions such as snow and night influence the performance. Pose estimators seem less sensitive to fog and rain.

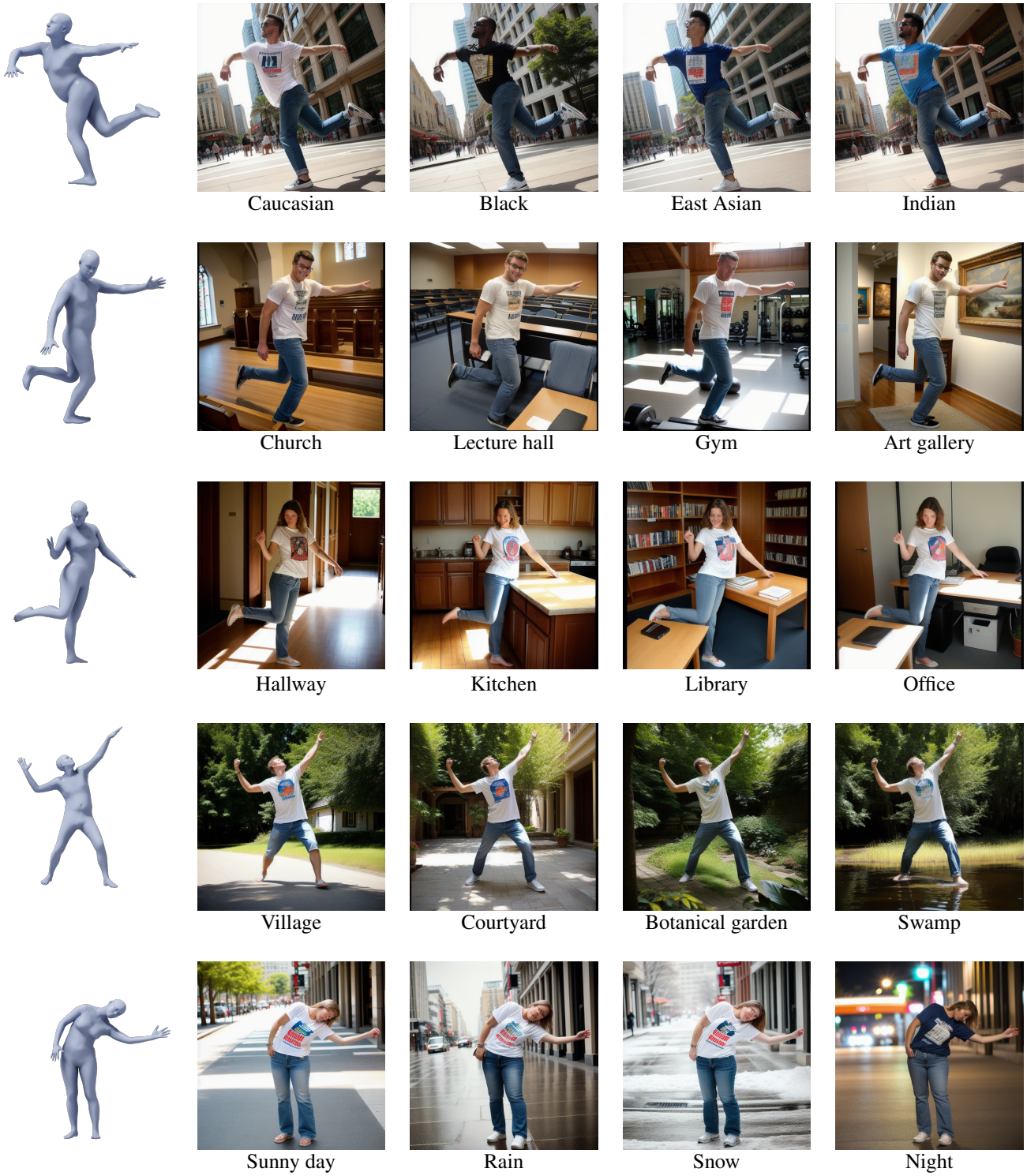

Figure S8. **Generated images for various attributes.** We present more examples of generated images. Based on a given ground truth pose (leftmost column), we generate images of people with different ethnicities, in different locations, or during different weather/lighting conditions. We start from a base prompt "Photo, adult caucasian male/female wearing a t-shirt in the city center at daytime." and modify a single attribute, e.g., "city center" to "gym". Images in the same row use the same initial noise for generation.

Figure S9. **Our method CN-3DPose generates diverse images.** Given a single pose and multiple prompts we generate images with CN-Pose, CN-Depth, CN-Multi and CN-3DPose. CN-Depth and CN-Multi fail to follow the prompt and generate flat backgrounds or the wrong clothing item (notice the absence of a coat for CN-Depth and CN-Multi in row 1). Each image is generated from a different randomly sampled noise.