

# Are Pose Estimators Ready for the Open World? STAGE: Synthetic Data Generation Toolkit for Auditing 3D Human Pose Estimators

Nikita Kister<sup>1</sup> István Sárándi<sup>2</sup> Anna Khoreva<sup>3</sup> Gerard Pons-Moll<sup>2,4</sup>

<sup>1</sup>Bosch IoC Lab, University of Tübingen <sup>2</sup>Tübingen AI Center, University of Tübingen

<sup>3</sup>Bosch Center for Artificial Intelligence

<sup>4</sup>Max Planck Institute for Informatics, Saarland Informatics Campus

{nikita.kister, istvan.sarandi, gerard.pons-moll}@uni-tuebingen.de anna.khoreva@bosch.com

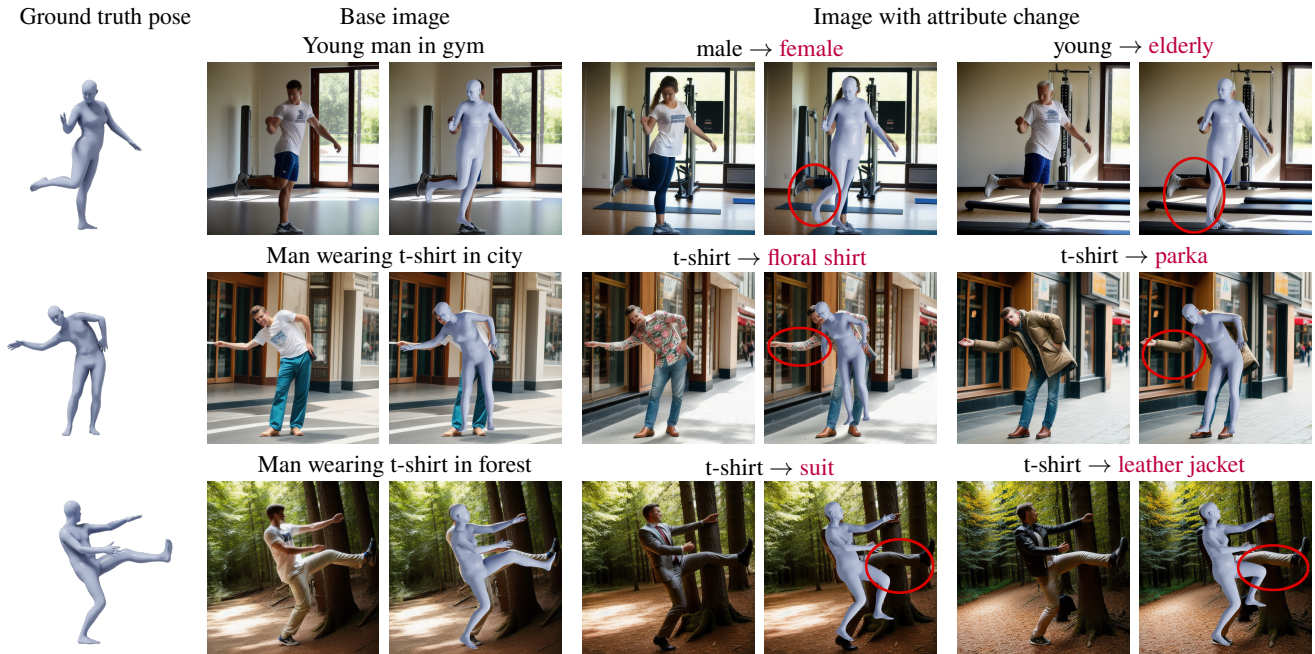


Figure 1. **How do 3D human pose estimators perform in the presence of various personal or environmental attributes (e.g., gender or clothing)?** Using text-to-image GenAI, our STAGE framework can answer such questions by dynamically generating custom benchmarks, targeting any user-specified attributes. By generating image pairs (base image and attribute-changed image), we can measure the pose estimator’s performance gap, enabling customizable fine-grained analysis unsupported by existing benchmarks.

## Abstract

The estimation of 3D human poses from images has progressed tremendously over the last few years as measured on standard benchmarks. However, performance in the open world remains underexplored, as current benchmarks cannot capture its full extent. Especially in safety-critical systems, it is crucial that 3D pose estimators are audited before deployment, and their sensitivity towards single factors or attributes occurring in the operational domain is thoroughly examined. Nevertheless, we currently lack a benchmark that would enable such fine-grained analysis. We thus present STAGE, a GenAI data toolkit for auditing 3D human pose

estimators. We enable a text-to-image model to control the 3D human body pose in the generated image. This allows us to create customized annotated data covering a wide range of open-world attributes. We leverage STAGE and generate a series of benchmarks to audit the sensitivity of popular pose estimators towards attributes such as gender, ethnicity, age, clothing, location, and weather. Our results show that the presence of such naturally occurring attributes can cause severe degradation in the performance of pose estimators and leads us to question if they are ready for open-world deployment. Code, data, and models will be released at <https://virtualhumans.mpi-inf.mpg.de/stage>.

## 1. Introduction

3D human pose estimation (HPE) from monocular images has been a main area of research in computer vision for over 30 years, and we have witnessed tremendous progress in recent years. HPE is a critical component of applications in collaborative robotics, systems that interact with or operate among humans, making it crucial to ensure its reliability. However, especially for safety critical systems such as autonomous vehicles and robots, it is crucial that the deployed models are not only accurate on average but also *robust towards elements naturally occurring in their operational domain*. A foremost precondition to ensuring robustness is the existence of rigorous and thorough evaluation protocols to stress-test these systems. Controlled experiments are crucial for a systematic analysis, isolating the influence of various factors or *attributes* one at a time, allowing more reliable conclusions about the causes of errors. The need for such evaluations has also been recently recognized by legislators in proposals such as the EU AI Act [13] and the UL 4600 standard by the American National Standards Institute [40].

Benchmarks consisting of real photographic data (Human3.6M [31] and 3DPW [74]) cannot provide such control – it is infeasible to capture a sufficient variety of people in the same exact pose while varying attributes. Furthermore, such benchmarks are laborious to capture, and attributes of interest vary depending on application. Computer graphics generated synthetic benchmarks [4, 6, 9, 57, 72, 76] provide a high degree of control over pose, camera location, clothing, *etc.* However, to achieve photorealism and diversity, they require large amounts of high-quality 3D assets. Clothing items have to be designed and physically simulated [6], and locations have to be modeled or captured. In addition, effective use of modern graphics pipelines requires expertise that HPE practitioners may lack. Overall, the current tools do not allow for a thorough auditing of HPE models as is required for their application in the open world.

We hence introduce STAGE, a new method to generate customizable HPE benchmarks, with which it becomes possible to measure the influence of individual attributes on HPE performance for the first time. For this, we build on recent progress in large-scale text-to-image models [61, 63, 65, 80], which have the potential to generate the required data quickly with an easy-to-use text-based user interface.

STAGE works as depicted in Fig. 1. Given a target attribute, *e.g.* a clothing item such as “parka”, we generate two sets of images, one where people wear parka coats and one where they do not, with a strategy to keep the appearance of the images roughly equal in other aspects. The performance difference between these two synthetic sets indicates the sensitivity of the pose estimator towards the attribute. While image synthesis models introduce a real-to-synthetic distribution shift, including potential artifacts, this only has limited impact on the sensitivity score. This is

because we always compare between two sets of generated images, which differ only in the target attribute, and the sets of images are generated at scale for many different poses. By enabling the generation of tailor-made benchmarks, STAGE allows researchers and practitioners to faithfully audit their models for their target application domains according to any attributes expressible in free-form text.

Creating a synthetic benchmark presents two main challenges: the generated images should 1) be sufficiently realistic and 2) contain cues coherent with the underlying 3D pose. None of the current generation methods satisfies such requirements. At best, recent methods allow control over 2D pose [33, 54, 83], but this is insufficient, as several different 3D poses can project to the same 2D poses [67, 70]. Other methods based on depth control [83] result in a lack of diversity. We therefore devise a new human image generator that is based on ControlNet [83] but is conditioned on a 3D rendering of SMPL [49] to control 3D pose and is trained with a carefully designed strategy on a combination of 3D pose datasets (providing accurate pose conditioning) and 2D datasets (providing more image diversity). This model produces high-quality images that are also consistent with the conditioning 3D pose. To demonstrate this, we generate a digital replica of the public 3DPW [74] dataset and show that a SOTA pose estimator’s performance on the synthetic replica is comparable to its performance on the real 3DPW.

STAGE can shed new light on aspects of current SOTA model performance, which can not be evaluated in current benchmarks. To demonstrate these capabilities, we use STAGE to extensively evaluate the sensitivity of popular SOTA HPE methods [6, 10, 24, 36, 38, 69, 82] to attribute categories such as clothing, location, weather, and protected attributes (gender, age, *etc.*), and make several interesting findings. For example, the best-performing methods, on average, are not always the most robust to attributes. In summary, our contributions are threefold.

- We propose STAGE, a customizable GenAI benchmark generator, along with a systematic evaluation protocol that allows, for the first time, to conduct controlled experiments to audit 3D pose estimators.
- Using STAGE, we empirically evaluate popular pose estimators on their sensitivity against attributes such as gender, age, and clothing, making findings that were not possible before.
- We build a text-to-image generative model that is able to create images that are coherent in terms of 3D pose.

We will release our code and data for future research.

## 2. Related Work

**3D human pose estimation benchmarks.** High-quality benchmarks have served as a major driving force behind progress in HPE. Real-world benchmarks [30, 31, 34, 74, 84,



85] are desirable but expensive to create at scale. They are either restricted to a studio [31, 85] or have a limited number of subjects [30, 34, 74, 84]. High variation of clothing, locations, gender, and ethnicity has not been possible to achieve with real data. Therefore, simulated benchmarks [4, 6, 9, 57, 72, 76] have risen in popularity, allowing the generation of accurately annotated data at scale. However, they require laborious asset design (clothes, 3D environments) as well as expertise to operate complex graphics, animation, and physics simulation pipelines. In practice, therefore even computer-graphics-based synthetic benchmarks are limited regarding the factors that can impact performance. In contrast, we argue that the *text input* of text-to-image models is a more flexible and easy-to-use interface for custom benchmark generation and control of attributes like clothing, locations, *etc.*

**Controllable human image generation** methods can be grouped into control given 2D pose [8, 33, 48, 54, 56, 83] and control given 3D pose [5, 11, 18, 25, 27, 28, 35, 39, 45, 55, 77, 79]. 2D pose conditioned methods such as ControlNet [83], T2I [54], HumanSD [33] and HyperHuman [48] extend StableDiffusion (SD) [63] with control via 2D human keypoints. They all exhibit good text control but do not allow precise control of human pose in 3D. While ControlNet is able to combine multiple control signals (such as depth and 2D pose) in a training-free procedure, the resulting model achieves only limited 3D controllability, as we will show in Sec. 4.3. For 3D pose control methods, we can broadly speak of GAN-based and Stable Diffusion-based methods. GAN-based methods [5, 18, 28, 55, 77] learn to generate 3D humans from 2D single-view image collections. Through techniques such as inverse skinning, they achieve excellent pose control. However, they generate people without backgrounds and do not offer text control. Methods based on SD [11, 39, 45, 58] or CLIP [27, 60, 79] offer textual control. However, for the task of benchmark creation, the image quality is often insufficient, and the generation process can take hours.

**Auditing with generative models.** The utility of generative models for benchmark creation and model auditing has been recognized before for the tasks of object classification and detection [1, 2, 7, 20, 23, 42, 44, 50, 53, 59, 64, 73, 75]. Some of these methods search for a semantically coherent subspace of images that result in failure [7, 20, 23, 53, 75], while others create counterfactual images to measure classifier robustness [1, 2, 50, 59, 73]. ImageNet-E [42] is a synthetic benchmark to evaluate attributes such as size, pose, and background. These methods typically rely on the availability of diverse benchmarks such as ImageNet [16] or on large held-out test datasets, which are not available for HPE. Additionally, unlike these classification and detection works, we are targeting HPE, a more fine-grained, high-dimensional regression problem requiring a more elaborate approach.

### 3. Method

We introduce a synthetic data generation toolkit named STAGE, which enables the creation of custom benchmarks for auditing HPE. We begin with a problem formalization in Sec. 3.1 and introduce our 3D pose conditioned image synthesis method in Sec. 3.2. In Sec. 3.3, we further describe the dataset generation pipeline, and in Sec. 3.4, we introduce the evaluation protocol to assess the sensitivity of pose estimators towards certain attributes.

#### 3.1. Evaluation with Controlled Generated Test Sets

A 3D pose estimator is a function  $f : \mathbf{x} \rightarrow \hat{\mathbf{P}}$  that maps an image  $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$  depicting a person to their corresponding 3D body pose  $\hat{\mathbf{P}} \in \mathbb{R}^{J \times 3}$ , represented as  $J$  3D joint locations. Typically, pose estimators are deployed in some operational domain where image  $\mathbf{x}$  and ground truth pose  $\mathbf{P}$  follow a distribution  $p(\mathbf{x}, \mathbf{P})$  and are evaluated with a metric  $L$  by computing the risk

$$R_f = \mathbb{E}_{p(\mathbf{x}, \mathbf{P})} [L(f(\mathbf{x}), \mathbf{P})]. \quad (1)$$

Sampling from  $p(\mathbf{x}, \mathbf{P})$  is commonly done by capturing real data, which is challenging to obtain in scale. This makes it difficult to evaluate the robustness of estimators to specific attributes in a controlled manner. Computer graphics generated synthetic data is an alternative but it is limited in realism and diversity due to the lack of large amounts of high-quality 3D assets available. Instead, we use generative models trained on Internet-scale data to sample from  $p(\mathbf{x}, \mathbf{P}) = p(\mathbf{x}|\mathbf{P})p(\mathbf{P})$ . To sample from  $p(\mathbf{x}|\mathbf{P})$ , we develop a 3D pose-conditioned diffusion model, and for  $p(\mathbf{P})$ , we sample poses from AMASS [52] or Motion-X [46].

We further want control over specific scene attributes such as clothing, location, *etc.*, which we summarize in the latent  $\mathbf{z}$ . Thus, we aim to sample from  $p(\mathbf{x}, \mathbf{P}|\mathbf{z}) = p(\mathbf{x}|\mathbf{P}, \mathbf{z})p(\mathbf{P}|\mathbf{z})$ . For the purpose of evaluation, we assume that a person can adopt the same body pose regardless of scene attributes, *i.e.*,  $p(\mathbf{P}|\mathbf{z}) = p(\mathbf{P})$ . Leveraging recent developments in text-to-image (T2I) generation, we use a T2I diffusion model to approximate  $p(\mathbf{x}|\mathbf{P}, \mathbf{z}) \approx p(\mathbf{x}|\mathbf{P}, t, \mathbf{n})$ , where  $t$  is a text prompt and  $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  is the initial noise for the denoising process. The  $\mathbf{z}$  is represented by  $t$  and  $\mathbf{n}$ . The text prompt allows us to generate images containing the required attributes, *e.g.*, “Photo, a Caucasian elderly male wearing a t-shirt and pants in the city center during daytime”. The noise vector encodes the rest of the details, such as the background layout and a specific appearance of the human not contained in the text prompt. We note that with diffusion sampling algorithms, such as DDIM [68], the generation process is deterministic given the initial noise  $\mathbf{n}$ . Thus sampling from  $p(\mathbf{x}|\mathbf{P}, t, \mathbf{n})$  amounts to first sampling the random noise vector  $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and then applying our diffusion model  $\mathbf{x} = g(\mathbf{P}, t, \mathbf{n})$ .

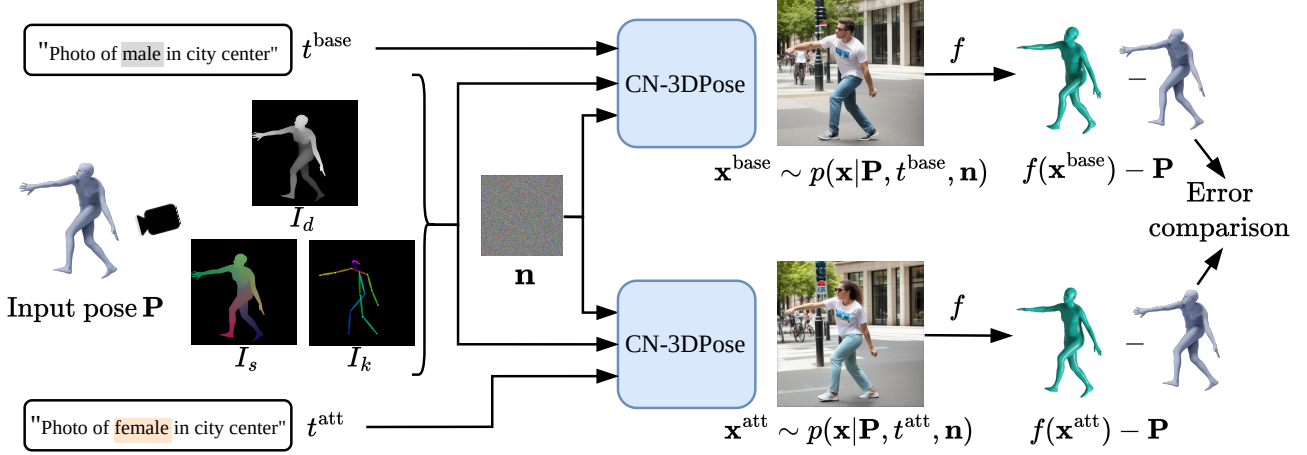


Figure 2. **Evaluation of estimators with STAGE** We start by generating a pair of human images given text prompts  $t^{\text{base}}$  and  $t^{\text{att}}$  and a desired 3D pose  $\mathbf{P}$  as a SMPL mesh. We render the posed SMPL mesh as a depth map, a dense semantic encoding and a 2D skeleton. These are input to our CN-3DPose to achieve 3D body pose control, while the text prompt control the overall appearance including the injected attributes to be tested. We use the same initial noise to generate both images to preserve visual similarity between the images. The result is an image pair that differ only by our attribute of interest. We apply the pose estimator to be tested on the pair, compute the prediction error and further process it to estimate the sensitivity of the pose estimator towards the attribute.

For controlled evaluation, we compare performance between generated images that differ only by our attribute of interest. We achieve this by using text prompts  $t^{\text{base}}$  and  $t^{\text{att}}$  that only differ by a single attribute description and using the same noise to generate both images. Note that  $t^{\text{base}}$  is constructed to describe a generic, most commonly occurring human appearance and scene, which in principle should not present a challenge for the pose estimator. We refer to the resulting sets of images as *base set* and *attribute set*. To compare the performance, we choose an evaluation function (Sec. 3.4) and estimate the risk

$$R_f(t^{\text{att}}) = \mathbb{E}_{\mathbf{n}, \mathbf{P}} \left[ L \left( f(g(\mathbf{P}, t^{\text{att}}, \mathbf{n})), f(g(\mathbf{P}, t^{\text{base}}, \mathbf{n})), \mathbf{P} \right) \right]. \quad (2)$$

In contrast to (1), we compare the performance of estimator  $f$  for synthetic base  $\mathbf{x}^{\text{base}} = g(\mathbf{P}, t^{\text{base}}, \mathbf{n})$  and attribute images  $\mathbf{x}^{\text{att}} = g(\mathbf{P}, t^{\text{att}}, \mathbf{n})$  on the average. Using pairs of synthetic images reduces the estimation noise caused by generation flaws. Computing the risk over a large number of pairs reduces the estimation noise caused by random factors. In Fig. S1 we show that we use a sufficient amount of samples.

### 3.2. Pose-Conditioned Image Generation

To condition, we render a depth map from the SMPL body model in a desired pose  $\theta \in \text{SO}(3)^{24}$  and given shape. The model  $g(\mathbf{P}, \mathbf{n}, t)$  has to meet two requirements. First, it must provide accurate *3D pose alignment* between the condition and the generated images. Second, the effect of the conditioning pose and text prompt has to be disentangled. Meaning the pose should not influence their shape or clothing. Our starting point to realize  $g(\mathbf{P}, \mathbf{n}, t)$  is ControlNet (CN) [83]. Zhang *et al.* [83] provide two models referred to as CN-Pose

and CN-Depth. However, these models do not meet the desired requirements.

CN-Pose [83] generates images conditioned on the 2D human pose. This model does offer reasonable image quality but suffers from poor 3D pose alignment since multiple 3D poses can project to the same 2D pose. Generated images can have swapped left and right limbs or generate back views instead of front views. CN-Depth [83] is conditioned on the given depth image, thus retaining the depth information and providing a better 3D pose alignment. However, using body pose depth leads to low diversity as the backgrounds are bland and clothing is mostly body-tight. This is because CN-Depth was trained with full image depth estimated from web images. Providing full image depth would restrict us to a finite amount of depth images, lowering the diversity. Combining CN-Pose and CN-Depth [83] results in an improved 3D pose alignment, but also negatively affects the diversity as can be observed in Tab. 2 and most obviously by visual inspection, see Fig. 10 and Fig. S9 in Supp. Mat. We refer to this combination as CN-Multi. None of these solutions fully meet our requirements.

**CN-3DPose.** Our image generation method, CN-3DPose, builds on the pre-trained CN-Depth and three key insights, which effectively leverage a combination of 3D and 2D pose datasets. First, we condition on isolated human body depth maps  $I_d$ , as opposed to real depth maps that include the background. This forces the model to use the text control and the initial noise to create diverse backgrounds and appearances while being coherent with the 3D pose. To fine-tune for this new task, we leverage existing public 3D pose datasets (see Sec. 4.1). Second, to avoid confusion between



Figure 3. **Images generated via STAGE.** We are able to generate images of people with different body shapes and appearances and in different locations, well-aligned with the given 3D ground truth pose (leftmost column). We use a base prompt “Photo, adult caucasian male/female wearing a t-shirt in the city center at day time sunny day” and modify a single attribute, *e.g.* “t-shirt” to “trench coat”.

body parts, we use an additional dense semantic conditioning  $I_s$ : a rendering of SMPL with each vertex colored according to its position in a canonical A-pose. Third, to mitigate over-fitting to the 3D pose dataset backgrounds and their limited variation in human appearance and poses, we also fine-tune using larger and diverse 2D pose datasets. Here, we simply condition on both 2D pose  $I_k$  and 3D pose, but zero out the 3D pose conditioning inputs ( $I_d$  and  $I_s$ ) when it is not available. This new model meets our requirements for diversity and 3D pose alignment. Overall, it provides us with a way to sample from  $p(\mathbf{x}|\mathbf{P}, \mathbf{z}) \approx p(\mathbf{x}|t, I_d, I_s, I_k, \mathbf{n})$ . Our conditioning inputs and the key steps of our method are depicted in Fig. 2.

### 3.3. Dataset Generation

**Set of attributes.** To evaluate the capabilities of pose estimators in the open world, we aim to cover several

groups of attributes. We cover attributes like clothing and location, as well as sensitive personal attributes such as ethnicity, age, and gender to detect potential biases in pose estimation methods. Finally, we also consider adverse conditions such as night, snow, or rain as it is a topic recognized in the autonomous driving community [14, 17, 51, 66], but not yet in the HPE community. This leads to a high-level scene description captured in the following prompt template: Photo, {ethnicity} {age} {gender} wearing {clothing} in {location} at {lighting condition} {weather}. For the base set construction, we populate the template to represent the average scene, with the intention not to present an extra challenge to the pose estimator, such as “Photo, caucasian young male wearing a t-shirt in the city center at daytime sunny day”. Note that the template structure can also be adapted by the user depending on the target operational domain of the pose estimator.



Pose estimator	Base error (mm) ↓		Sensitivity to attributes (PDP, %) ↓							
	MPJPE	PA-MPJPE	Location (outdoor)	Location (indoor)	Fairness	Clothing	Weather	Texture	Mean	
SPIN [38]	122.50	90.27	15.79	19.65	12.93	29.13	17.43	29.26	20.70	
PARE [36]	118.81	88.98	13.34	15.84	9.36	21.66	12.97	18.43	15.27	
MeTRAbs [69]	89.80	67.77	17.06	19.57	11.31	22.93	15.88	16.94	17.28	
PyMAF-X [69]	115.81	84.03	9.83	12.49	5.40	14.61	8.51	10.60	10.24	
HMR 2.0 [24]	102.40	75.21	9.31	12.41	5.34	15.76	7.57	10.75	10.19	
BEDL-CLIFF[6]	113.14	84.22	17.30	20.77	15.61	23.36	15.47	19.42	18.65	
SMPLer-X [10]	S32	117.84	90.61	15.62	19.54	11.20	21.76	15.88	18.51	17.08
	B32	107.00	80.40	13.30	17.24	7.81	17.38	12.57	13.13	13.57
	L32	101.83	75.98	14.48	17.84	7.73	16.53	12.30	12.40	13.54
	H32	104.79	76.00	14.35	18.55	6.75	15.11	10.54	10.74	12.67

Table 1. **Evaluating sensitivity of various pose estimators with STAGE.** We evaluate the percentage of degraded poses and contrast it against the general performance on the base set. We find that appearance attributes such as clothing and texture are the most impactful. In addition, pose estimators are more sensitive to indoor locations compared to outdoor locations.

**Variation reduction.** To reduce factors of variation between the base and attribute sets, we fully specify the template, even for attributes not directly relevant to the experiment. For example, not specifying the daytime can lead to a mix of images that depict daytime and nighttime. Additionally, the noise sharing scheme described in Sec. 3.1 reduces the variation between base and attribute image and thus the variance of our estimation of the risk Eq. (2). The effect can be observed in Fig. 3.

**Quality control.** To deal with noise and errors in the synthesis process, we introduce a filtering mechanism. First, the target attribute should be present in the image. Second, the person should exhibit our desired pose. To ensure the attribute presence, we employ a Visual Question Answering (VQA) based filtering. Based on the input prompt, we construct questions about the presence of the attributes, following TIFA [29]. The questions, jointly with the image, are fed into the VQA model BLIP2 [41]. If the answer of the VQA model regarding the attribute is negative, the image is discarded. To ensure consistency between image and 3D pose, we apply filtering based on 2D human pose alignment. (Filtering using a 3D pose method would risk removing the challenging and interesting images.) Hence, we discard an image if OpenPose’s [12] 2D keypoints predictions in the generated image deviate from the projected 3D keypoints by more than a given threshold. Since we jointly generate the base and attribute images, we discard the whole pair if at least one of them is discarded.

### 3.4. Evaluation Protocol

We aim to identify cases where the prediction is *degraded* compared to the base set in the presence of the target attribute. We consider an estimated pose degraded if any joint is significantly displaced. Hence, we use the Maximum Joint

Error (MaxJE) – the highest joint error of a predicted pose. For a ground truth pose  $\mathbf{P} \in \mathbb{R}^{J \times 3}$ , and predicted pose  $\hat{\mathbf{P}} \in \mathbb{R}^{J \times 3}$  (Procrustes-aligned) with joints  $\mathbf{p}_i, \hat{\mathbf{p}}_i \in \mathbb{R}^3$  respectively, the maximum joint error  $\text{MaxJE}(\hat{\mathbf{P}}, \mathbf{P}) = \max_{1 \leq i \leq J} \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|$ . We employ Procrustes alignment to account for differences in camera angle between the pair of images. A predicted pose  $\hat{\mathbf{P}}_{\text{att}} = f(\mathbf{x}^{\text{att}})$  is then called degraded if its error is worse than the base case by a larger margin than a given threshold  $\tau$ , i.e., when

$$L(\hat{\mathbf{P}}_{\text{att}}, \hat{\mathbf{P}}_{\text{base}}, \mathbf{P}, \tau) = [\text{MaxJE}(\hat{\mathbf{P}}_{\text{att}}, \mathbf{P}) - \text{MaxJE}(\hat{\mathbf{P}}_{\text{base}}, \mathbf{P}) > \tau] \quad (3)$$

is one. The Percentage of Degraded Poses (PDP) for a threshold  $\tau$  for an attribute with dataset size  $N$  is then

$$\text{PDP} = \frac{1}{N} \sum_{i=1}^N L(\hat{\mathbf{P}}_{\text{att},i}, \hat{\mathbf{P}}_{\text{base},i}, \mathbf{P}_i, \tau), \quad (4)$$

which is exactly the estimate for the risk Eq. (2) when using the degradation criterion Eq. (3) as the loss  $L$  in Eq. (2). Note that the PDP is nonnegative by construction. It aims to count the especially high-risk instances where a pose estimator performs well under a common benchmark scenario but degrades when moving to the open world, indicating non-robustness. To compute the PDP for a category, we average the PDP for each attribute in the category. The overall PDP score for a pose estimator is the mean across all categories.

## 4. Experiments

This section is divided into three parts. First, we provide implementation details. Second, we use STAGE to evaluate the sensitivity of popular pose estimators towards certain attributes. Third, we evaluate the quality of the images generated by our GenAI toolkit STAGE.



Figure 4. **State-of-the-art estimators can break with a simple texture change.** We use the base prompt “Photo, caucasian adult male/female wearing a shirt in the city center during day time.” and modify “shirt” to “floral shirt”(left image) and “checkered shirt”(right image).

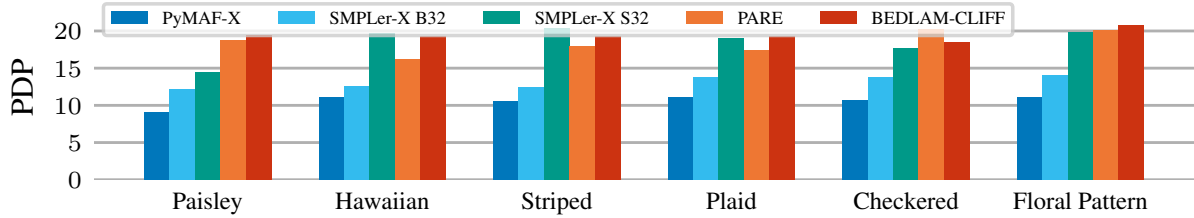


Figure 5. **Pose estimators are susceptible to texture changes.** Clothing texture can have a large impact on performance of pose estimators, e.g. on PARE. The impact of each texture is roughly equal.

#### 4.1. Implementation Details

**Training.** We initialize our CN-3DPose with weights from a pre-trained ControlNet-Depth [83]. We train it using AGORA [57], HUMBI [78, 81], SHHQ [22], COCO [47] and a set of human scans [3, 62, 71]. The remaining technical details can be found in the Supp. Mat.

**Pose estimators.** We evaluate, MeTRAbs [69], HMR 2.0 [24], PyMAF-X [82], SMPLer-X [10] and BEDLAM-CLIFF [6, 43] as these are the current state-of-the-art estimators. Note BEDLAM-CLIFF achieves state-of-the-art performance by training only on the synthetic datasets AGORA and BEDLAM [6, 57]. Here, we examine how effective this approach is in generalizing to the open world. In addition, we test the popular estimators SPIN [38] and PARE [36].

**Data generation.** We use a set of 1500 poses for generation. For experiments on 3DPW, we use farthest point sampling following [15] to obtain diverse poses. For experiments on AMASS, we use the subset provided by [15] and subsample it to 1500. For AMASS, we balance the gender and select 750 male and 750 female labeled poses. For our pose quality filter, we select a threshold of 50 px (see Sec. 3.3), as we observed that it is a good trade-off between accuracy and generation speed. We generate an image for a pose at most 13 times and check the quality filter. If none of the 13 generated images passes the quality filter, the pose is discarded (details on this are in the Supp. Mat.). Therefore, datasets for different attributes can have different poses. To control for this, we compare results between different attributes always on the intersection of valid poses.

#### 4.2. Examining Sensitivity to Various Attributes

In the following, we apply STAGE to examine the sensitivity of pose estimators with respect to different open-world attributes. We emphasize that our attribute selection is to showcase the method, but users can choose an entirely different set of attributes – code for generation and evaluation will be made public. The prompts used to create these datasets are provided in the Supp. Mat. Fig. 3 shows a sample of our generated images. Notice how we can generate diverse images in a controlled manner. We can keep the overall image appearance fixed while changing one specific attribute at a time – an ideal controlled scenario that is unfeasible with purely real data. We use poses from AMASS to generate the data. For evaluation of sensitivity, we use the (MaxJE and PDP with  $\tau = 50$  mm) metrics defined in Sec. 3.4.

**Summary.** Overall, we observe a significant percentage of degraded poses when testing for a specific attribute as summarized in Tab. 1. The attributes that target the body appearance directly (Clothing and Texture) have the most significant effect and can lead to large prediction errors (see Figs. 4 and 6). Full results are available in the Supp. Mat.

**Diverse training data reduces sensitivity.** A natural question is how training data and model size affect the sensitivity of the HPE methods. Our results indicate that a combination of large-scale data with a large model size is crucial for robustness. This can be observed from the SMPLer-X model family. Each variant uses a different-sized ViT [37] backbone. While all models were trained on the same amount of data, only large models use the data effectively, leading to a reduc-



Figure 6. **State-of-the-art estimators can break with a simple clothing change.** We use the base prompt “Photo, caucasian adult male/female wearing a shirt in the city center during day time.” and modify “shirt” to “parka” (left image) and “trench coat” (right image).

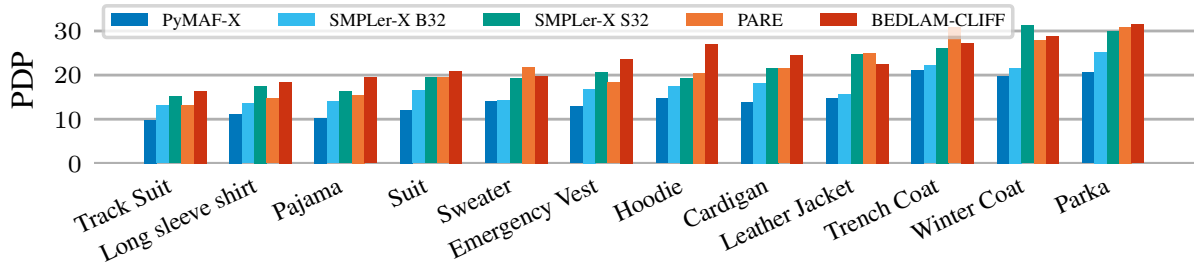


Figure 7. **Clothing impacts performance.** Clothing has the largest impact on performance. Especially items that cover most of the body such as coats, can impact the performance in a negative way.

tion of PDP across attributes. This is most pronounced for attributes of fairness, clothing, and clothing texture attributes, *i.e.*, attributes that affect the appearance of the person. HMR 2.0 is a piece of further evidence for that, as it was trained on a large-scale dataset with a ViT-H backbone. Conversely, SPIN and PARE were trained with the least amount of data and are the most sensitive estimators (largest PDP). However, diversity alone is not enough. BEDLAM-CLIFF (SOTA on real benchmarks) was trained on purely synthetic data, designed to be highly diverse, but is very sensitive to attribute changes across the board. This further evidences that our STAGE method sheds new light on the performance of SOTA methods in the open world, which can not be measured in existing benchmarks.

**Architectural choices matter.** While large-scale datasets lead to lower PDP, the performance of PyMAF-X indicates that smart architectural choices can be an efficient way to achieve robustness as well. PyMAF-X was trained on roughly the same datasets as PARE but is less sensitive to attribute changes. We attribute this to the iterative feedback loop to align the prediction to the input image in PyMAF-X. While the general performance is only about 3 mm better than PARE, it results in a significantly more robust method, which again was impossible to verify before our STAGE method.

**Clothing significantly affects pose estimation.** While intuitive, for the first time, we empirically verify that clothing significantly affects pose estimation performance. Figs. 5 and 7 show the sensitivity to each clothing item and to

different texture patterns, respectively. We observe that large body covering items such as coats and long jackets are most impactful, leading to a degradation of up to 30% of the predictions and over 15% across all estimators. In addition, a simple change such as the pattern of a shirt can lead to up to 20% of degraded predictions. Further, Fig. 5 suggests that each pattern is equally impactful.

**Location and weather.** It is commonly accepted in the HPE community that outdoor scenes are more challenging than indoor scenes [32]. However, our results show that SOTA methods degrade more when the attribute is indoor; see Tab. 1. We attribute this to the fact that standard benchmarks captured indoors are typically lab settings with very simple backgrounds, whereas our generated images have more realistic backgrounds, including potential occlusions. This hints at the fact that indoor HPE in the open world is not necessarily easier than outdoors. Perhaps unsurprisingly, bad weather conditions, in particular snow, result in the most degradation of poses.

**Pose estimators are robust against protected attributes.** Fig. 9 shows the sensitivity to protected attributes – attributes that should not be used to make decisions. Overall, current pose estimators are less sensitive towards protected attributes (such as ethnicity) compared to clothing and location. This might be due to the fact that these attributes usually only affect a small part of the body, making them less important overall. We note, however, that gender does seem to affect predictions more (see also Fig. 8) than other attributes –





Figure 8. **State-of-the-art estimators can break with a gender or age change.** We use the base prompt “Photo, caucasian adult male wearing a shirt in the city center during day time.” and modify “male” to “female”(left image) and “adult” to “elderly”(right image).

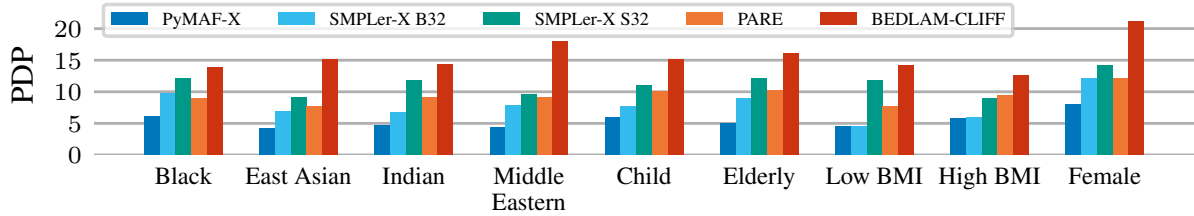


Figure 9. **Fairness analysis.** We consider multiple attributes related to fairness in computer vision. A subset is presented here. Pose estimators seem to be influenced by gender and age attributes while being robust against different ethnicities.

which indicates that to make these methods fair, more work is needed. One can also observe that BEDLAM-CLIFF is quite sensitive to every attribute despite being designed to cover a range of skin tones. This indicates that computer graphics generated data might not be sufficient to achieve good open-world performance.

### 4.3. Quality Assessment of STAGE Data

In this section we evaluate the quality of our generated images. We use our CN-3DPose model to construct a synthetic replica of a HPE benchmark. We do not aim for an exact copy but for a similar attribute distribution, *e.g.*, same clothing, gender, and location. This way, the dominant distribution shift comes from the different source of images. We then compare two things: 1) pose estimation performance difference between real and synthetic data, which we refer to as *pose gap*, and 2) the FID [26] score to measure how close our replica is to the real-world dataset in terms of visual quality and diversity.

We use 3DPW [74] for this experiment because it is well-established and covers different real-world scenes such as parks, city centers, and restaurants. To capture the attributes within an image for our prompt template, we use a VQA model [41]. See Supp. Mat. for details. As image synthesis baselines, we use CN-Pose and CN-Multi, and we use the current SOTA HPE method (MeTRAbs-ACAE [69]) on 3DPW to compute the pose gap.

We present the results in Tab. 2. CN-Pose is the weakest generator in terms of pose alignment as the pose gap 32.83 mm is the largest. Only 2D pose conditioning is not sufficient due to the aforementioned loss of 3D information

(see Sec. 3.3). In addition, it is prone to swap left and right limbs, as can be observed in Fig. 10 (left and right legs are swapped). In contrast, CN-Multi is able to generate images with accurate 3D pose alignment, achieving a smaller pose gap of 9.19 mm. However, its diversity is limited, as indicated by the FID score of 60.8. In addition, it often does not follow the text prompt, creates flat backgrounds, and is biased towards generating body-tight clothing, limiting diversity even further. Both phenomena can be observed in visual examples in Fig. 10 (more examples can be found in the Supp. Mat.). This makes pose estimation easier on the images and can be seen as good pose alignment, but at the expense of diversity, which is not what we want.

CN-3DPose is able to achieve high diversity and accurate 3D pose alignment thanks to our proposed design choices explained in Sec. 3.2. While the pose gap is slightly higher 13.11 vs 9.19 of CN-Multi, we attribute this to our higher diversity of backgrounds and clothing, making pose estimation harder. This can be clearly seen visually (see Supp. Mat.) and is also reflected in our much lower FID score 60.8 vs 39.7. We further investigate the reason for a pose gap of 13.11 mm. We hypothesize that there are two factors of influence. First, the backgrounds of our generated images are more complex and diverse compared to 3DPW, making the estimation task harder. Second, we observe artifacts in faces, feet, and hands in our generated images, which might influence estimation performance. Thus, we compare estimation performance again after blurring faces, feet, and hands, after background removal, and both. Results are in Tab. S1. After applying the blur and removing the background, the pose gap shrinks



Figure 10. **Our method CN-3DPose generates diverse images with good 3D pose alignment.** Given a single pose and multiple prompts we generate images with CN-Pose, CN-Multi and CN-3DPose. CN-Multi fails to follow the prompt and generates flat backgrounds or the wrong clothing item (notice the flat wall for CN-Multi in row 2). CN-Pose fails to follow the pose and generates the feet consistently in the wrong order. Only CN-3DPose (Ours) is able to generate diverse images while offering good 3D pose alignment.

Generator	Pose Gap ↓	FID ↓	Pose Accuracy	Diversity
CN-Pose	32.83	58.8	✗	✗
CN-Multi	9.19	60.8	✓	✗
CN-3DPose	13.11	39.7	✓	✓

Table 2. **Quantifying the domain gap.** We compare CN-3DPose with CN-Pose and CN-Multi. “Pose alignment” denotes the MeTRAbs pose estimator’s performance gap (in PA-MPJPE) between the real 3DPW images and our synthetic replica. CN-3DPose generates diverse, high-quality images with good 3D pose alignment.

to 6.5 mm, suggesting that more complex backgrounds and lower quality of faces, feet, and hands are causing the gap. However, we compute the PDP for our experiments between sets of generated images, base and attribute set. Thus, any flaw would be present in both sets, therefore not influencing the conclusion. In addition, our main contribution is the evaluation framework and it will benefit from any future improvement of generative models [19, 21]. We conclude our proposed CN-3DPose is sufficiently accurate and diverse to generate images for controlled auditing of HPE methods.

## 5. Conclusion

3D human pose estimation has been a main area of research in computer vision for over 30 years, and the community has seen a lot of progress in terms of generalization and good benchmarks. However, are these methods robust to attributes such as clothing, weather, or gender? For the first time, we can answer these questions empirically, shedding new light on pose estimation methods with STAGE, a method to build custom benchmarks for 3D human pose estimation at no cost. We built upon text-to-image models and adapted their capabilities to generate diverse, realistic images with 3D human pose control. This allows us to vary one attribute at a time and do controlled experiments. We use the model to create synthetic benchmarks, allowing one to audit their pose estimators for specific operational domains. We use our method to create controlled benchmarks, and test current SOTA pose estimators against attributes such as clothing, background texture, weather, and fairness. We make several interesting findings with our method: most estimators are sensitive to clothing (intuitive but never verified empirically) and texture. To a lesser degree, most methods are also affected by certain protected attributes such as ethnicity, gender, or age.

Overall, the methods are sensitive to several attributes; this is concerning and requires further investigation to make pose estimation methods fair and safe. Future work will investigate generating more complex scenes, including multiple people and control over objects. We will release code and data to allow researchers and practitioners to evaluate their 3D pose estimation methods against their desired attributes, allowing a much deeper understanding of how they will perform in the target domain.

## Acknowledgments

We thank Riccardo Marin for proofreading and the whole RVH team for the support. Nikita Kister was supported by Bosch Industry on Campus Lab at the University of Tübingen. Nikita Kister thanks the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program for support. István Sáránci and Gerard Pons-Moll were supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 409792180 (Emmy Noether Programme, project: Real Virtual Humans). GPM is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645 and is supported by the Carl Zeiss Foundation.

## References

- [1] Maximilian Augustin, Valentin Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. In *NeurIPS*, 2022. 3
- [2] Maximilian Augustin, Yannic Neuhäus, and Matthias Hein. Analyzing and explaining image classifiers via diffusion guidance. In *CVPR*, 2024. 3
- [3] AXZY Dataset. Axyz dataset. <https://secure.axyz-design.com>. Accessed: 2024-03-07. 7, 15
- [4] 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*, 2022. 2, 3
- [5] Alexander W. Bergman, Petr Kellnhofer, Yifan Wang, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *NeurIPS*, 2022. 3
- [6] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 2, 3, 6, 7
- [7] Valentin Boreiko, Matthias Hein, and Jan Hendrik Metzen. Identifying systematic errors in object detectors with the SCROD pipeline. In *ICCV Workshops*, 2023. 3
- [8] Tim Brooks and Alexei A. Efros. Hallucinating pose-compatible scenes. In *ECCV*, 2022. 3
- [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. *arXiv:2110.07588*, 2021. 2, 3
- [10] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, 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 *NeurIPS: Datasets and Benchmarks*, 2023. 2, 6, 7
- [11] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. DreamAvatar: Text-and-shape guided 3D human avatar generation via diffusion models. In *CVPR*, 2024. 3
- [12] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *TPAMI*, 2019. 6, 15
- [13] Council of European Union. Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts no 2021/0106 (cod). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>, 2021. 2
- [14] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2018. 5
- [15] Delmas, Ginger and Weinzaepfel, Philippe and Lucas, Thomas and Moreno-Noguer, Francesc and Rogez, Grégory. PoseScript: 3D human poses from natural language. In *ECCV*, 2022. 7, 16
- [16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 3
- [17] Shuai Di, Qi Feng, Chun-Guang Li, Mei Zhang, Honggang Zhang, Semir Elezovikj, Chiu C. Tan, and Haibin Ling. Rainy night scene understanding with near scene semantic adaptation. *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, 22(3):1594–1602, 2021. 5
- [18] Zijian Dong, Xu Chen, Jinlong Yang, Michael J. Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to generate 3D avatars from 2D image collections. In *ICCV*, 2023. 3
- [19] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scal-



- ing rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 10
- [20] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. In *ICLR*, 2022. 3
- [21] Flux Model. Flux model. <https://blackforestlabs.ai/>, 2020. Accessed: 2024-07-16. 10
- [22] 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. 7, 15
- [23] Irena Gao, Gabriel Ilharco, Scott Lundberg, and Marco Tulio Ribeiro. Adaptive testing of computer vision models. In *ICCV*, 2023. 3
- [24] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 6, 7
- [25] Artur Grigorev, Karim Isakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. StylePeople: A generative model of fullbody human avatars. In *CVPR*, 2021. 3
- [26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a nash equilibrium. In *NIPS*, 2017. 9
- [27] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhonggang Cai, Lei Yang, and Ziwei Liu. AvatarCLIP: Zero-shot text-driven generation and animation of 3D avatars. *TOG*, 41(4):1–19, 2022. 3
- [28] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. EVA3D: Compositional 3D human generation from 2D image collections. In *ICLR*, 2023. 3
- [29] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023. 6
- [30] 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 *CVPR*, 2022. 2, 3
- [31] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 2, 3
- [32] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *3DV*, 2020. 8
- [33] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. HumanSD: A native skeleton-guided diffusion model for human image generation. In *ICCV*, 2023. 2, 3
- [34] 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 *ICCV*, 2023. 2, 3
- [35] Markus Knoche, István Sárándi, and Bastian Leibe. Reposing humans by warping 3D features. In *CVPR Workshops*, 2020. 3
- [36] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 2, 6, 7
- [37] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 7
- [38] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2, 6, 7
- [39] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. DreamHuman: Animatable 3D avatars from text. In *NeurIPS*, 2023. 3
- [40] Underwriters Laboratories. UI 4600 evaluation of autonomous products, 2023. 2
- [41] 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. 6, 9, 15, 16
- [42] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. ImageNet-E: Benchmarking neural network robustness via attribute editing. In *CVPR*, 2023. 3
- [43] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 7
- [44] Hao Liang, Pietro Perona, and Guha Balakrishnan. Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation. In *ICCV*, 2023. 3
- [45] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxing Tang, Yangyi Huang, Justus Thies, and Michael J.

- Black. TADA! text to animatable digital avatars. In *3DV*, 2024. 3
- [46] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X: A large-scale 3D expressive whole-body human motion dataset. In *NeurIPS: Datasets and Benchmarks*, 2023. 3
- [47] 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. 7, 15
- [48] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skokhodov, Dahua Lin, Xihui Liu, Ziwei Liu, Sergey Tulyakov, and Yanyu Li. HyperHuman: Hyper-realistic human generation with latent structural diffusion. In *ICLR*, 2024. 3
- [49] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2
- [50] Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De la Torre. Zero-shot model diagnosis. In *CVPR*, 2023. 3
- [51] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 5
- [52] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 3
- [53] Jan Hendrik Metzen, Robin Huttmacher, N. Grace Hua, Valentyn Boreiko, and Dan Zhang. Identification of systematic errors of image classifiers on rare subgroups. In *ICCV*, 2023. 3
- [54] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 2, 3
- [55] Atsuhiko Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *ECCV*, 2022. 3
- [56] Mirela Ostrek, Soubhik Sanyal, Carol O’Sullivan, Michael J. Black, and Justus Thies. Environment-Specific People. *arXiv:2312.14579*, 2023. 3
- [57] 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 *CVPR*, 2021. 2, 3, 7, 15
- [58] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *arXiv:2209.14988*, 2022. 3
- [59] Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. LANCE: Stress-testing visual models by generating language-guided counterfactual images. In *NeurIPS*, 2023. 3
- [60] 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. 3, 15
- [61] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125*, 2022. 2
- [62] RenderPeople Dataset. RenderPeople dataset. <https://renderpeople.com>, 2020. Accessed: 2024-03-07. 7, 15
- [63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 15
- [64] Nataniel Ruiz, Barry-John Theobald, Anurag Ranjan, Ahmed Hussein Abdelaziz, and Nicholas Apostoloff. MorphGAN: One-shot face synthesis GAN for detecting recognition bias. In *BMVC*, 2021. 3
- [65] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2
- [66] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *ECCV*, 2018. 5
- [67] Cristian Sminchisescu. 3D human motion analysis in monocular video: Techniques and challenges. In *Human Motion: Understanding, Modelling, Capture, and Animation*, pages 185–211. Springer Netherlands, 2008. 2
- [68] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3
- [69] István Szárándi, Alexander Hermans, and Bastian Leibe. Learning 3D human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *WACV*, 2023. 2, 6, 7, 9
- [70] Camillo J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *CVIU*, 80(3):349–363, 2000. 2

- [71] Twindom Dataset. Twindom dataset. <https://web.twindom.com>. Accessed: 2024-03-07. 7, 15
- [72] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2, 3
- [73] Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv:2302.07865*, 2023. 3
- [74] 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 *ECCV*, 2018. 2, 3, 9
- [75] Olivia Wiles, Isabela Albuquerque, and Sven Gowal. Discovering bugs in vision models using off-the-shelf image generation and captioning. In *NeurIPS Workshops*, 2022. 3
- [76] 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 *ICCV*, 2023. 2, 3
- [77] Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 3DHumanGAN: 3D-aware human image generation with 3D pose mapping. In *ICCV*, 2023. 3
- [78] Jae Shin Yoon, Zhixuan Yu, Jaesik Park, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions and benchmark challenge. *TPAMI*, 45(1):623–640, 2023. 7
- [79] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. CLIP-Actor: Text-driven recommendation and stylization for animating Human meshes. In *ECCV*, 2022. 3
- [80] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. In *ICLR*, 2024. 2
- [81] 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. 7, 15
- [82] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: Towards well-aligned full-body model regression from monocular images. *TPAMI*, 45(10), 2023. 2, 7
- [83] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 4, 7, 15
- [84] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. EgoBody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, 2022. 2, 3
- [85] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 3



# Are Pose Estimators Ready for the Open World? STAGE: Synthetic Data Generation Toolkit for Auditing 3D Human Pose Estimators

## Supplementary Material

### A. Further Analysis on Real vs. Synthetic Gap

Here, we analyze in detail the reasons behind the error gap between 3DPW and its synthetic replica that we reported in Tab. 2 of the main paper. We hypothesize that the differences between synthetic and real images are largely concentrated on the background, faces, hands and feet. To verify this, we conduct experiments with blurring faces, hands and feet, and background removal in both real and generated images following the protocol from Sec. 4.3. As the results in Tab. S1 show, jointly these almost cancel the MPJPE gap, showing the body is recognizable for the pose estimator to a similar degree, validating the overall quality of generated images. The quality of faces and limbs is an aspect that is actively being worked on in the image generation community, therefore we can expect future models to further reduce the gap. Background removal improves the performance on synthetic images more than on real ones; this suggests that our synthetic backgrounds are more complex than those in 3DPW, enabling testing under more challenging conditions.

### B. Implementation Details

#### B.1. Model Training

The CN-3DPose architecture is based on [83]. We adapt the input layer to take nine channel dimension, and we initialize it with weights from a pre-trained CN-Depth [83]. We train our model using AGORA [57], HUMBI [81], SHHQ [22], COCO [47] and a set of human scans [3, 62, 71]. For the image datasets, we create training images by center-cropping around each person in the image.

**Preprocessing.** In our evaluation, we do not consider occlusions, therefore we remove images where the person is significantly occluded. Specifically, for COCO, we drop images where less than ten keypoints are visible, and for AGORA, we project the SMPL mesh vertices onto the image and drop it if more than 20% are outside the image. We resize images to  $512 \times 512$ . The cropping can result in the image having black regions due to the required padding. We mask these regions during training to prevent the model learning from generating them. Our model is a latent diffusion model [63]. Therefore, we resize the mask to  $64 \times 64$  (spatial dimension of the latent) and apply it to the denoising loss during training. For our evaluation, we only consider single-person images. Hence, when possible, we mask other people in the image during training. COCO provides instance and crowd masks, and AGORA provides instance masks we

		PA-MPJPE ↓			MPJPE ↓		
Blurred face, hands, feet	Backgr. removed	Real	Synth.	Gap	Real	Synth.	Gap
		47.16	60.27	13.11	69.31	79.76	10.45
	✓	45.60	55.25	9.65	64.84	70.93	6.09
✓		49.56	60.30	10.74	70.90	79.81	8.91
✓	✓	47.46	53.96	<b>6.50</b>	68.52	69.41	<b>0.89</b>

Table S1. **Impact of background and certain body parts on the pose error gap.** The pose error gap between real and synthetic data reduces when we remove the background or blur faces, hands and feet, which are body parts that current image generators often struggle with.

can use. The person mask is combined with the black border mask.

**Caption generation.** For training CN-3DPose, we need image captions. We use the image captioning model BLIP2 [41] to create textual descriptions for each image. To avoid overfitting, we construct multiple captions for each image. First, we sample five captions with BLIP2 and filter them using CLIP [60] to ensure image-text alignment. During training, a random caption from the set is selected for the image. We are also using synthetic data for training (AGORA and the scans). Thus, we want to prevent the model from generating their synthetic look. We do so by adding the word “Rendering” at training time to captions of images from these datasets. Using negative prompting at inference time further prevents the model from creating images similar to AGORA and human scans.

**Training.** For the hyperparameters, we follow [83]. We are training with datasets of different sizes. To balance data, we sample from datasets with probability proportional to their size. We train our model on 4 NVIDIA A100 40GB for nine days.

#### B.2. 2D Pose Estimation Filter

To identify low-quality samples in the generation, we apply a 2D keypoint-based filter mechanism. We predict 2D keypoints for our generated images with OpenPose [12] and compare them against the projected ground truth keypoints. We are generating images based on the SMPL mesh. To get a 2D keypoint in the OpenPose skeleton format, we first apply a joint regressor on the SMPL mesh to get 3D joints in the OpenPose format. Then, we project them onto the image. OpenPose predicted keypoints have different variances depending on their type, e.g., wrists have lower

variance than hips. Therefore, we only consider wrist, ankle, shoulder, elbow, and knee keypoints for our filtering. We compute the 2D position error for each keypoint and compare it against a threshold (50px in our case). The image is invalid if the error of at least one selected keypoint is higher than the threshold. The image is also invalid if multiple people are detected in the image. Invalid images are dropped and not used for evaluation.

### B.3. Constructing the Synthetic Replica of 3DPW

We sample poses from 3DPW with farthest point sampling to create a diverse subset of poses. Each pose has a corresponding bounding box containing the target person that we use to create image crops. Thus, each pose has a corresponding image. The construction of the replica aims to recreate these images using STAGE. We require poses, which are already given, and text descriptions of the images. To create a textual description, we use our prompt template `Photo, caucasian {gender} wearing {clothing} in {location} during {weather} at {daytime}` and fill in the values for the attributes. The dataset was recorded during the day, and all participants were Caucasian. We take the gender from the annotations. To identify clothing and locations, we use a VQA model (BLIP2 [41]) and ask “What is the person in the foreground wearing?” and “Where is the person located?”.

In total, we use 1500 poses from 3DPW for the generation. We use the provided camera parameters to create the inputs for our model. Since we use pose-based and VQA-based quality filtering, we do not generate 1500 valid images, because we interrupt the generation process after 13 attempts. The number of valid images the models generated is

1. CN-Pose: 1021
2. CN-Multi: 1367
3. CN-3DPose: 1372.

For a fair comparison, we only selected poses that resulted in valid images for all methods. In total, we evaluate 981 poses.

### B.4. Generating Data for Attribute Experiments

**Sampling poses.** For our attribute sensitivity experiments, we used poses from AMASS. Specifically, we use the subset provided by PoseScript [15] and sample 1500 poses (750 male, 750 female) using furthest point sampling to create a set of diverse poses.

**Rendering model input.** AMASS does not provide camera parameters in contrast to 3DPW. We set the focal length to 4 by visual inspection. The camera is placed such that its center is aligned with the root joint of the pose. To have the whole body in the frame, we set the distance of the camera so that the body is barely visible in the frame. We set the rotation to the identity matrix and instead rotate the pose around its vertical axis. We want to avoid introducing variances that come from self-occluded joints. For this, we rotate the

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

Table S2. **Remaining number of poses used for evaluation.** Due to our filtering mechanism it is possible that we never generate a valid image pair for a given pose. These poses are dropped from the evaluation. It is also possible that different attributes have different set of valid poses. Therefore, we use only poses that are valid for all attribute in a given category when we compute our metrics. Here we present the number of valid poses for each attribute category, that we used to compute our metrics.

body so that its front is facing the camera, and wrists, ankles, shoulder, and elbow joints are visible in the camera.

**Choosing a base prompt.** For our experiments, we chose a base prompt that describes a scene that does not pose any challenges to the pose estimator. In general, our base prompt is the following `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 specific experiments, we deviate from it. For the location (indoor) experiment, we replace “city center” with “hallway” to provide a neutral indoor location. For the location (outdoor), we replace “city center” with “village” to evaluate “city center” itself. For the fairness experiments, only body shape and the gender experiment deviate from the base prompt. For gender, we use “male” for our base dataset and “female” for our attribute dataset. For the body shape experiment, we used “adult with average BMI” for our base dataset and “adult with low/high BMI” for the two attribute datasets. We observed that specifying “average BMI” is not required since there is little deviation from the average body shape.

**Generating images.** To generate the attribute datasets, we modify the corresponding base prompt by replacing the base attribute with the target attribute. For example, if we generate data to examine coats, we replace “t-shirt and pants” with “coat” to create the attribute prompt. Similar to the experiment on the synthetic 3DPW, we do not generate 1500 valid images because we use pose and VQA-based filtering. Therefore, we use only poses that result in valid images for all attributes in the category. For example, if we examine outdoor locations, we only use poses that result in valid images for all our chosen locations. We list the number of used poses per category in Tab. S2. One concern is that the number of poses is too low to estimate the PDP reliably. Therefore, we investigate how the estimate changes with different numbers of samples. To do so, we compute the PDP with a different number of poses. We show the results in Fig. S1. Overall, the PDP is stable after the first couple of hundred poses. Thus, we conclude that we have used a sufficient number of samples.

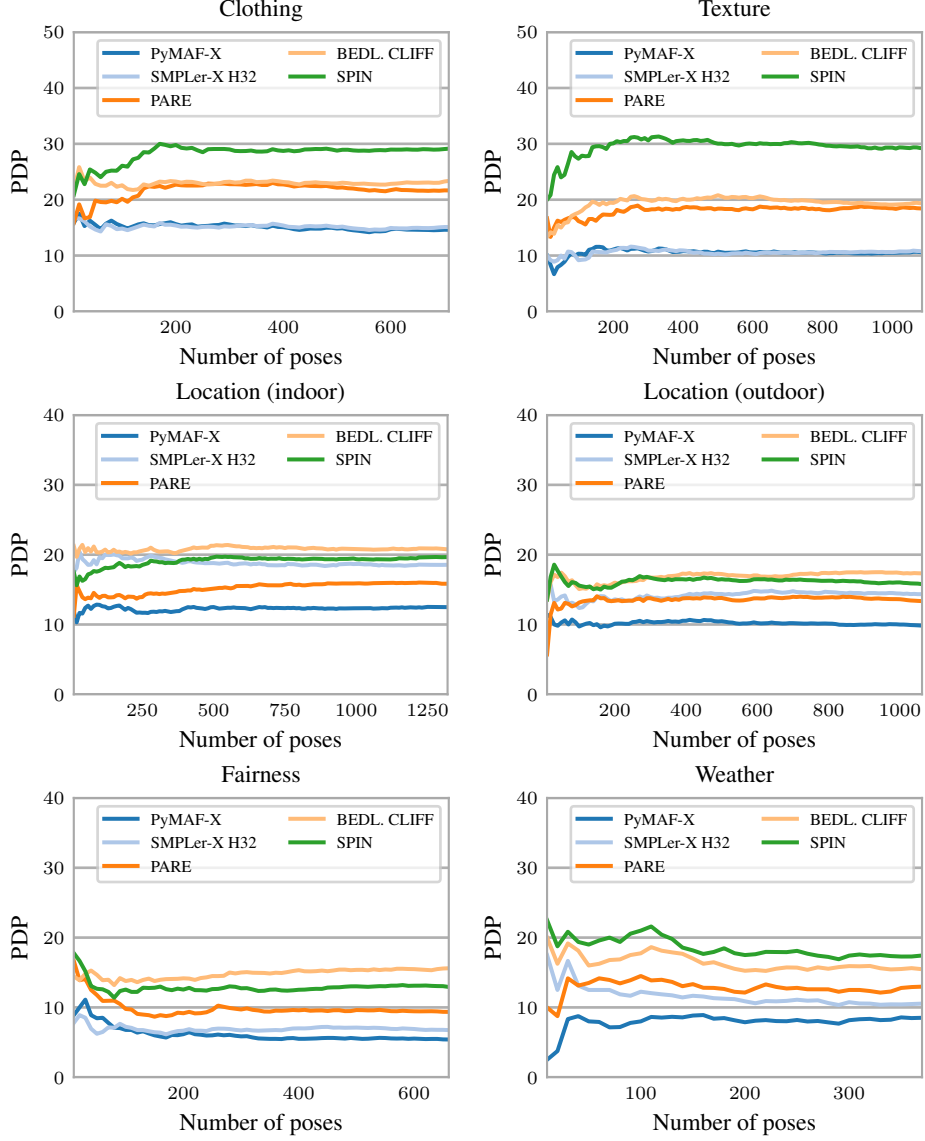


Figure S1. **Our number of samples is sufficient.** We plot how our results per category change based on the number of valid poses used to compute them. Overall the PDP remains stable after the first couple of hundred of poses.

## C. Full Results

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

Overall, 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 can be best observed in Figs. S2 and S4, where we see a continual increase of PDP across all estimators from left to right. This continual increase indicates that the more difficult attributes also affect the best models, even if only by a little.

For the outdoor location experiments, we observe that

each location has 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 show a clear increase in PDP from left to right. The most difficult indoor locations such as “bar”, “restaurant” or “kitchen” have in common that tables are a frequently occurring object that can introduce occlusions. Indeed, we observe that the images of these attributes can depict occluded body limbs, which suggests the leading cause of error for these attributes to be occlusions.



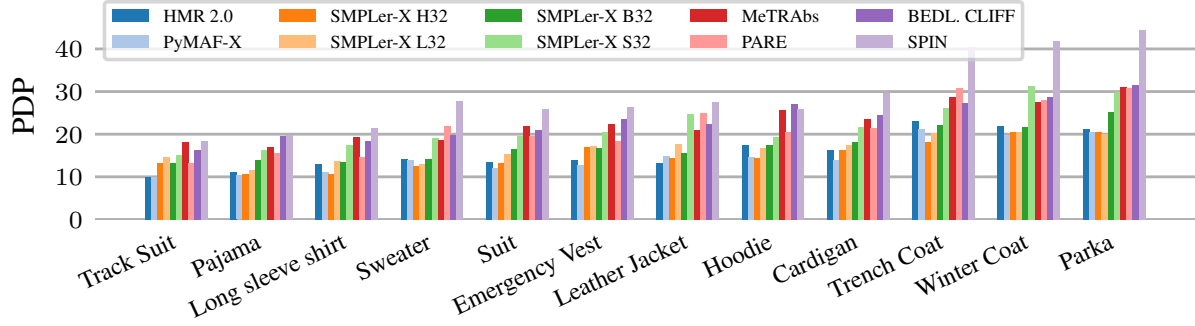


Figure S2. **Clothing impacts performance.** Clothing has the largest impact on performance. Especially items that cover most of the body, such as coats, can impact the performance in a negative way.

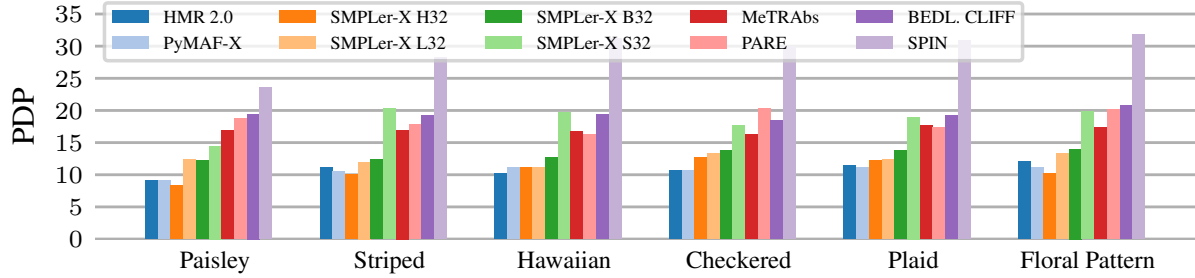


Figure S3. **Pose estimators are susceptible to texture changes.** All textures lead to about the same PDP, indicating that a texture change influences performance regardless of what that texture is. SPIN is particularly sensitive as up to 30 percent of the poses are degraded.

## 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.

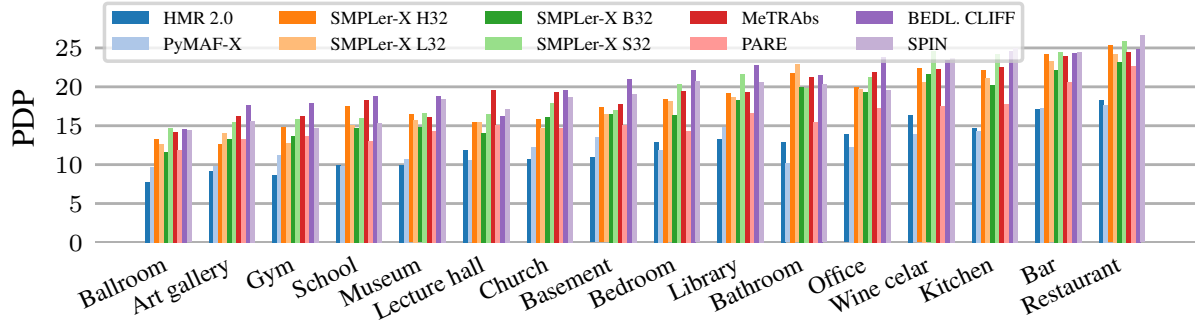


Figure S4. **Influence of indoor locations.** The continual increase of the PDP indicates that some locations are more challenging than others. Especially, “Restaurant”, “Bar” and “Wine cellar” have the most impact on performance.

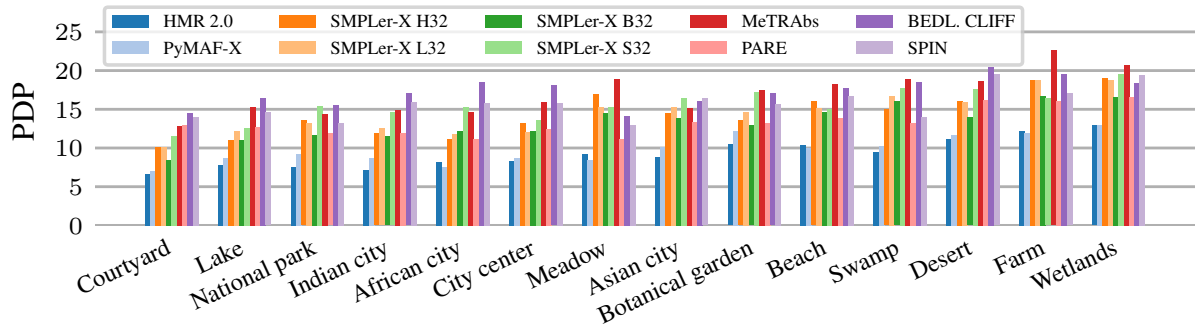


Figure S5. **Influence of outdoor locations.** Most outdoor locations have a similar impact on performance, “Swamp and “Wetlands” pose the most challenges to pose estimators.

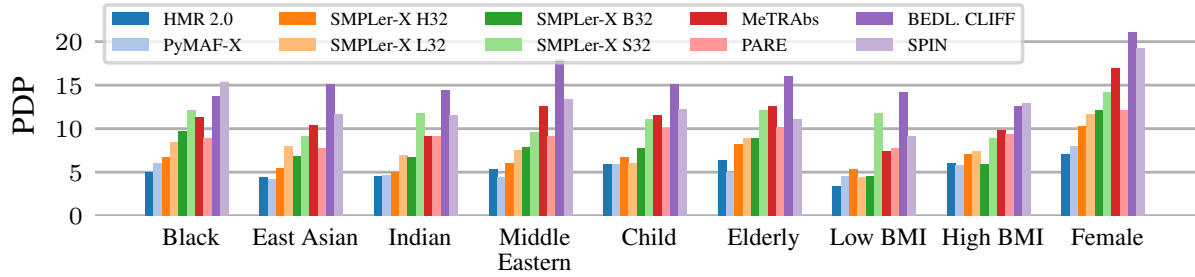


Figure S6. **Fairness analysis.** We consider multiple attributes related to fairness in computer vision. Pose estimators are robust against protected attributes. However, gender and age.

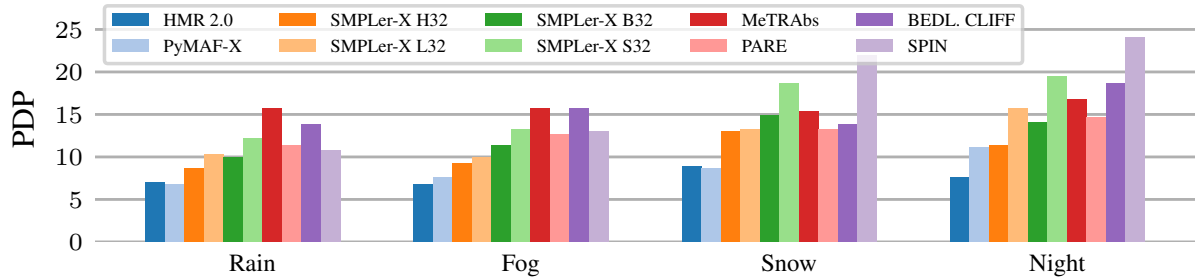


Figure S7. **Influence of adverse conditions.** Adverse conditions such as snow and night influence the performance. Pose estimators seem less sensitive to fog and rain.

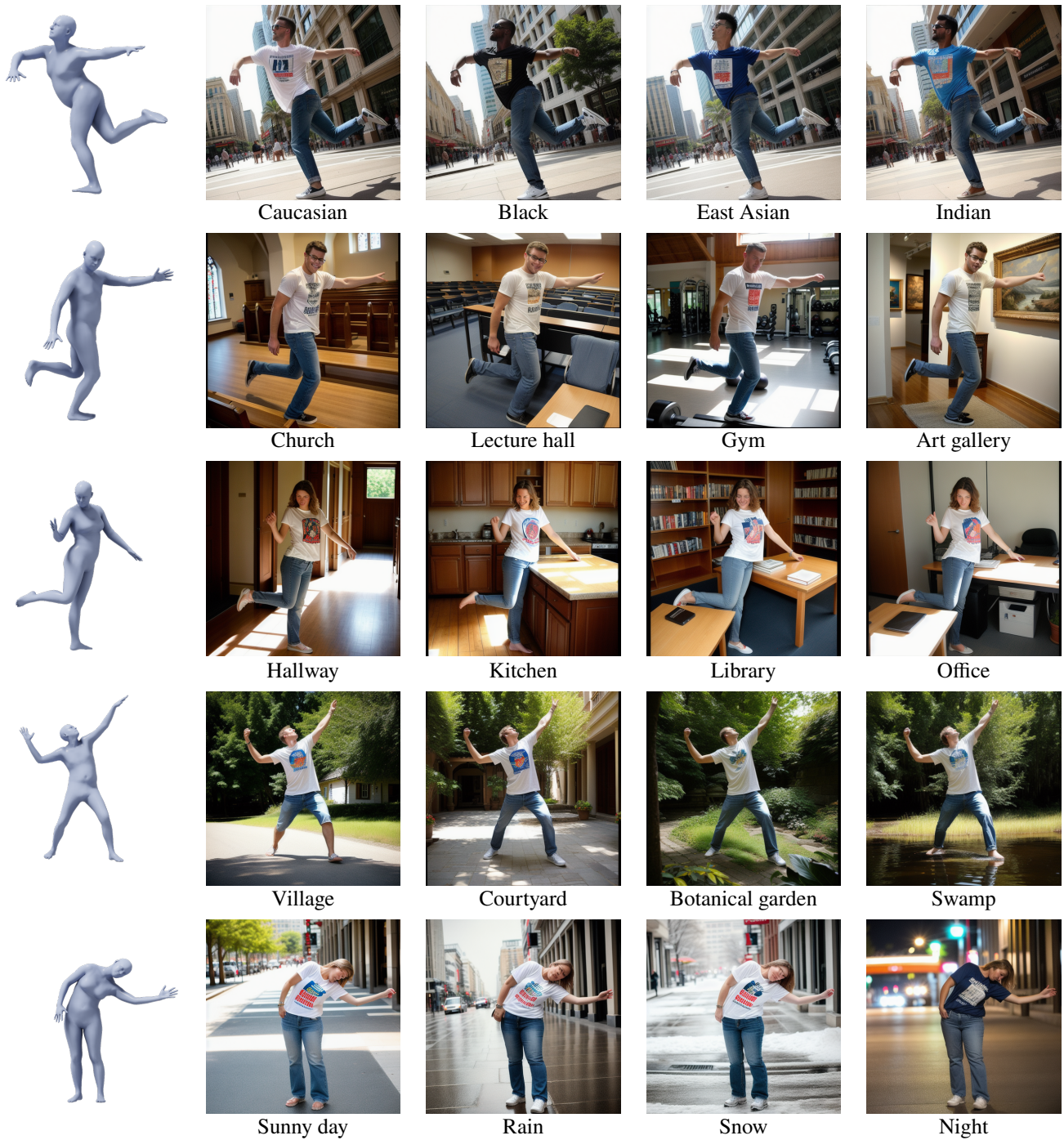


Figure S8. **Generated images for various attributes.** We present more examples of generated images. Based on a given ground truth pose (leftmost column), we generate images of people with different ethnicities, in different locations, or during different weather/lighting conditions. We start from a 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.



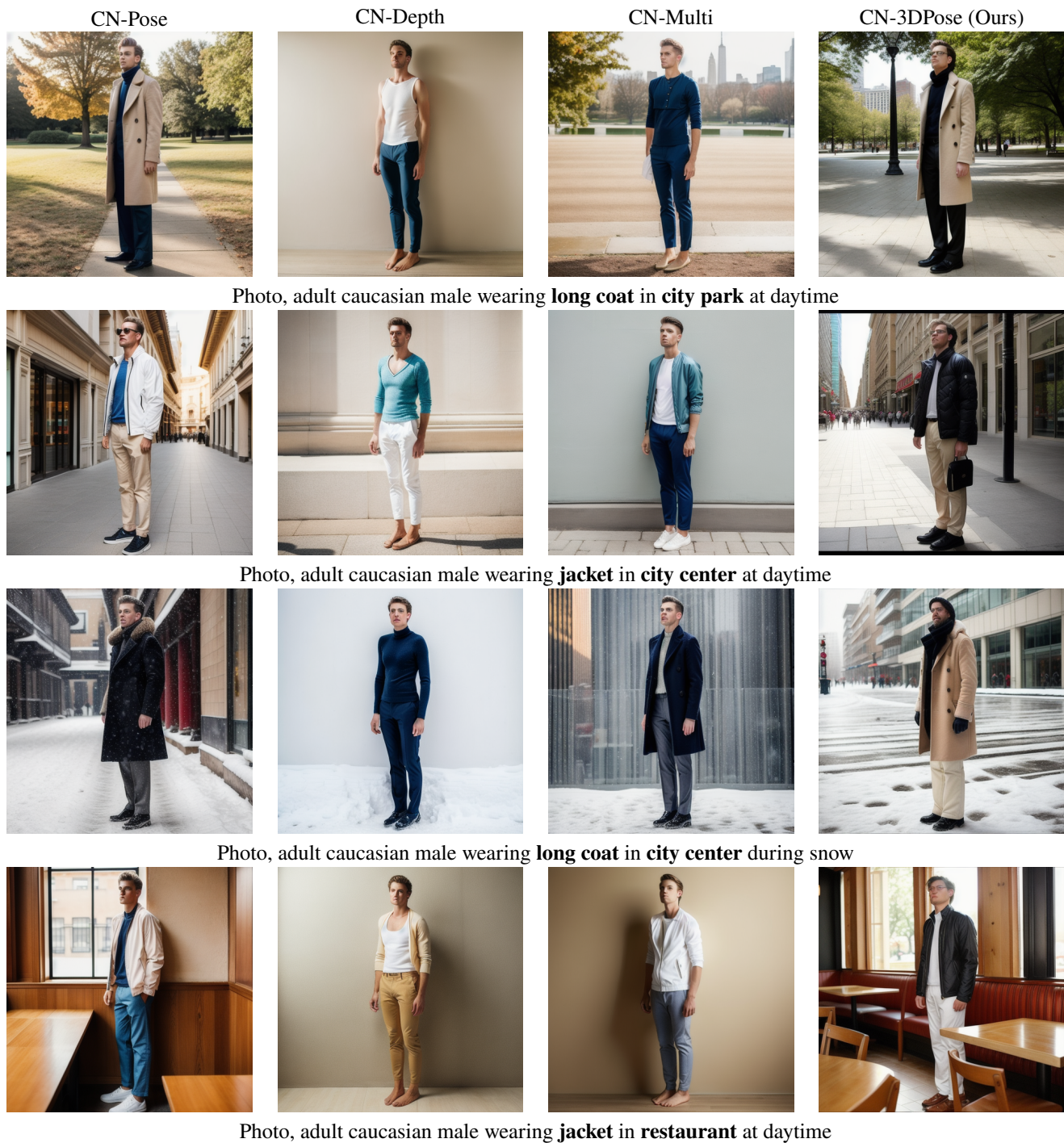


Figure S9. **Our method CN-3DPose generates diverse images.** Given a single pose and multiple prompts we generate images with CN-Pose, CN-Depth, CN-Multi and CN-3DPose. CN-Depth and CN-Multi fail to follow the prompt and generate flat backgrounds or the wrong clothing item (notice the absence of a coat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