Diffusion-HPC: Synthetic Data Generation for Human Mesh Recovery in Challenging Domains

Zhenzhen Weng, Laura Bravo-Sánchez, Serena Yeung-Levy
Stanford University
{zzweng, lmbravo, syyeung}@stanford.edu

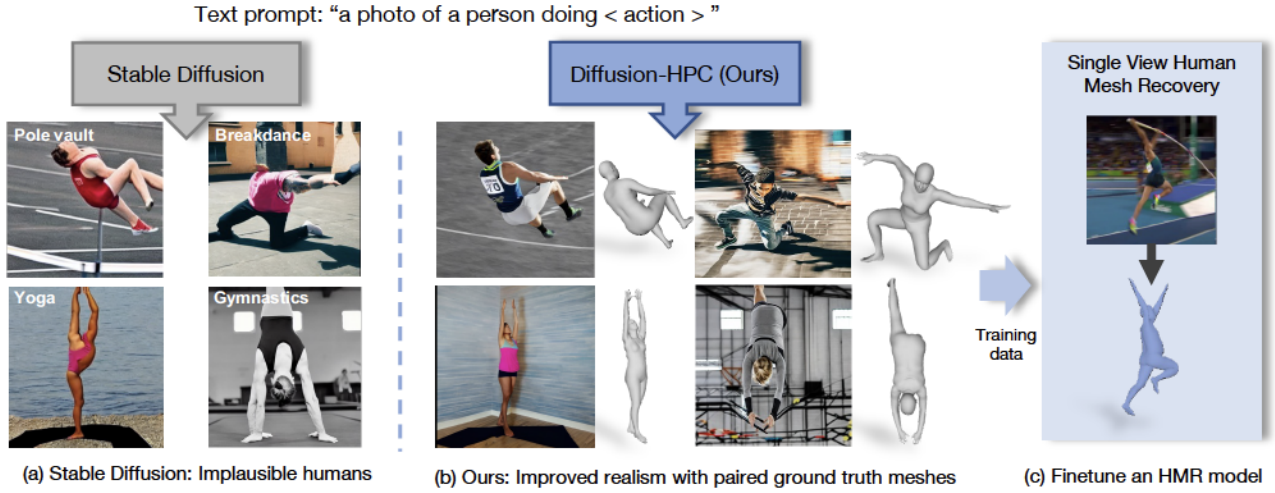

Figure 1. We propose Diffusion model with Human Pose Correction (Diffusion-HPC), a synthetic image generation strategy with paired with ground-truth meshes to improve the performance of Human Mesh Recovery (HMR) models on domains with challenging poses and/or limited data. Diffusion-HPC is a text-conditioned method that addresses the implausibility of human generations from Stable Diffusion [14], a large pre-trained text-conditioned generative model, while preserving the inherent flexibility of such models.

Abstract

Recent text-to-image generative models have exhibited remarkable abilities in generating high-fidelity and photo-realistic images. However, despite the visually impressive results, these models often struggle to preserve plausible human structure in the generations. Due to this reason, while generative models have shown promising results in aiding downstream image recognition tasks by generating large volumes of synthetic data, they are not suitable for improving downstream human pose perception and understanding. In this work, we propose a Diffusion model with Human Pose Correction (Diffusion-HPC), a text-conditioned method that generates photo-realistic images with plausible posed humans by injecting prior knowledge about human body structure. Our generated images are accompanied by 3D meshes that serve as ground truths for improving Human Mesh Recovery tasks, where a shortage of 3D training data has long been an issue. Furthermore, we show that Diffusion-HPC effectively improves the realism of human generations under varying conditioning strategies. [1]

1. Introduction

In recent years, large-scale text-conditioned image generation models such as GLIDE [35], Imagen [43] and Stable Diffusion [14] have impressed the research community with their exceptional generative and compositional capabilities, owing to their training on extremely large image-text datasets [44] and use of advanced model architectures [11, 21]. Not only do these generative models drastically elevate the quality and efficiency of content creation, but they also exhibit promising potential for enhancing other visual tasks. As shown in He et al. [18], large text-conditioned generative models such as GLIDE [35] are able to gener-

[1] Code: https://github.com/ZZWENG/Diffusion_HPC

ate high-quality images targeted for a specific label space (i.e. domain customization), thus making it an ideal choice for synthetic data generation to aid in downstream image recognition tasks such as single-view Human Mesh Recovery (HMR) [29] where securing annotations can be not only costly, but incompatible in the wild.

Despite the benefits of synthetic data for image recognition tasks, these text-conditioned generative models have thus far lacked the capability of advancing human pose understanding tasks. This is because these models do not explicitly model the underlying structure of human bodies and thus frequently encounter difficulties in preserving realistic human anatomy in their generated outputs. As Figure 1 (a) shows, generating realistic human poses embedded in plausible scenes is a known limitation [23] of generative diffusion models such as Stable Diffusion [14].

In this work, we present Diffusion model with Human Pose Correction (Diffusion-HPC), a method that addresses the implausibility of human generations from large pre-trained text-conditioned generative models. Our intuition is that we can rectify the generated unrealistic humans (e.g. with additional limbs in non-anatomical locations) by integrating stronger human pose priors within the generation process. Thereby, we extend the capability of pre-trained diffusion models, such as Stable Diffusion, to produce a large variety of synthetic scenes for a target domain with minimal user input. Further, unlike base diffusion models our approach produces pairs of images and ground truth meshes as a result of including body pose priors in the generation process. These image-mesh pairs can then be employed to improve existing single-view Human Mesh Recovery methods on challenging data-scarce domains (See Figure 1b, c). In summary, we make the following contributions.

- Motivated by the implausible humans produced by diffusion models, we propose a simple and effective method Diffusion-HPC to rectify the implausibility of human generations that often occur in Stable Diffusion [14] results. To the best of our knowledge, our work presents the first training-free method that addresses the challenges in generating realistic humans by injecting human body structure priors within the generation process.
- We show that the synthetic images with corresponding 3D ground truth produced by our method are capable of adapting Human Mesh Recovery models to challenging domains (e.g. competitive sports) where supervision is limited and hard to obtain. Models finetuned with Diffusion-HPC's synthetic data achieve 2.6% PCK and 4.6 PA-MPJPE improvement on SMART [9] and SkiPose 3D [41, 46], respectively.
- We quantitatively validate the improved quality of our generated images over existing text-to-image as well as state-of-the-art pose-to-image generative models.

2. Related Work

2.1. Using synthetic data to improve HMR

Previous works [18] have recognized the capability of state-of-the-art text-conditioned generative models [35] for generating training data for downstream image recognition tasks. However, the poor quality of the generated person images effectively precludes the extension of this capacity to tasks such as 3D human pose understanding (e.g. human mesh recovery). Due to the challenge in collecting 3D ground truths for end-to-end training of human mesh recovery models, many previous works have considered leveraging synthetic data. Typically, these works create 2D renderings of 3D posed human models from graphics engines [12, 37, 48], with Black et al. [6] being the most comprehensive effort. However, this approach possesses multiple disadvantages. First, the variety of the generated poses is limited by the pose data source. Second, a large and diverse training set is needed to cover all possible poses of interest, which makes storing and sharing such data costly and inefficient. To address these, recent work Weng et al. [51] proposes a data-efficient way by rendering SMPL bodies with poses sampled from the estimated pose distributions from real data, but since the body textures are predicted and warped from real images, the renderings are not photo-realistic. Analogously, Sengupta et al. [45] generates synthetic data online to improve diverse body shape estimation.

In contrast, using conditional generative models [14, 21] such as ours to synthesize data has a few advantages. First, large generative models can produce high-fidelity photo-realistic images closer to real data since they are trained on internet-scale real-world data (e.g. LAION-5B [44]). Second, they allow easy control of the generation style via detailed prompting, and stochasticity in the generation process results in more diverse and potentially unlimited synthetic data. However, although there have been some attempts to explore the use of generative models for image classification and object detection [18, 57], their usage in human mesh recovery has not yet been investigated due to the poor quality of the generated person images. Diffusion-HPC is the first approach that uses conditional generative models to produce synthetic data that are useful for human mesh recovery, broadening the range of downstream utilities of SoTA generative models.

2.2. Conditional generation of posed humans

Recently, there has been a growing focus on conditional generation of posed humans in the form of images/videos [22, 49, 54, 56], body models [10, 16] or NeRF [8, 13, 52]. For synthesizing posed humans, most works focus either on text-conditioned or pose-conditioned generation. In terms of text-conditioned approaches state-of-the-art general im-

age generation models such as Stable Diffusion [14] have shown impressive capability in producing high-resolution and realistic images. But as a known limitation [23], they frequently struggle to preserve the correct anatomy of human bodies. An alternate line of research focuses directly on text-conditioned human pose or motion generation [10, 16]. These generative models are trained on large 3D human motion database [33] with paired textual descriptions. But since they output parameters of human body models [32], they bypass the issue of preserving anatomical structure. However, since human motion databases do not come with paired RGB data, these works are unable to produce human textures and background.

On the other hand, previous works have explored generating images of full-body humans conditioning on body pose [1, 7, 27, 34]. AlBahar et al. [1], Knoche et al. [27], Men et al. [34] consider the task of "reposing", where the goal is to synthesize images of people in a novel pose, based on a reference image of that person and the new pose. More recently and relevant to our work, Brooks *et al*. [7] proposed a pose-conditioned image synthesis model that dispenses with the reference images by generating reasonable backgrounds. In contrast to the above works, our proposed Diffusion-HPC is a person image synthesis method flexible enough to allow for both text and pose conditioned generation and does not require additional training or explicit pose annotations from a target domain to produce diverse humans and scenes. Another closely related work is ControlNet [56], where posed-conditioned images are obtained from generative models, yet unlike our method ControlNet requires finetuning on large amounts of real paired data (i.e. 2D keypoints, images and captions).

2.3. Editing & composing large pre-trained models

Foundation models (e.g., Imagen [43], Stable Diffusion [14]) that are trained on large amounts of broad data have demonstrated impressive generative and few-shot learning capabilities across a wide spectrum of tasks. Additionally, the scale of information they have seen and learned, allows these models to be adapted to further downstream tasks. For these reasons, editing or composing large pre-trained models has been widely studied recently. Among these works the most closely related to our approach are the training-free methods that utilize pre-trained diffusion models to perform global or local image editing [3, 4, 19, 36] (e.g. inpainting, style transfer, etc.). They are "training-free" in the sense that editing is done by injecting knowledge into the denoising process during inference and therefore no additional model finetuning is needed. Analogously, our method Diffusion-HPC improves plausibility of human generations by injecting human body priors in the form of posed SMPL [32] body models.

3. Diffusion-HPC

3.1. Background

Latent diffusion models. Diffusion models are deep generative models that generate samples from a desired distribution by learning to reverse a gradual noising process. The sampling process starts from noise sampled from a standard normal distribution, which are refined into a series of less-noisy latents that eventually lead to the desired generation. For more details, please refer to Dhariwal *et al*. [11] and Ho *et al*. [21]. Latent Diffusion uses a perceptual compression model, a variational autoencoder (VAE) that projects the data distribution into a latent space, where the conditional diffusion process operates. Previous works [3] have shown that editing in the latent space is faster than pixel space editing [4] and helps to avoid pixel-level artifacts. Our method uses two latent diffusion models under the hood, a text-to-image model where the denoising is conditioned on the text input, and a depth-to-image model where the depth map is used as additional conditioning.

SMPL body model. We use the Skinned Multi-Person Linear (SMPL) model [32] to represent the 3D mesh of the human body. SMPL is a differentiable function $\mathcal{M}(\theta, \beta)$ that takes a pose parameter $\theta \in \mathbb{R}^{69}$ and shape parameter $\beta \in \mathbb{R}^{10}$, and returns the body mesh $\mathcal{M} \in \mathbb{R}^{6890\times 3}$ with 6890 vertices. The 3D joint locations $X \in \mathbb{R}^{k\times 3} = \mathcal{W}\mathcal{M}$ are regressed from the vertices, using a pre-trained linear regressor $\mathcal{W}$, where k is the number of joints.

3.2. Data generation process

Diffusion-HPC consists of three main steps. As a first step, we leverage a text-to-image model (i.e. Stable Diffusion [14]) to produce an initial generation of a posed person. Second, we predict the pose of the person and determine the difficulty level of the pose in the initial generation using a pose prior. We observe that Stable Diffusion tends to generate worse anatomy on more difficult poses. Thus, if the initial generation contains a hard pose, we render the depth map of the predicted body mesh taking into consideration the occlusion from other objects in the image. The depth map serves as the human structure prior. In the final step, we use the context information (in the form of image latents) from initial generation as a starting point, and leverage a depth-to-image model (i.e. a fine-tuned version of Stable Diffusion) to produce final generations by conditioning on the depth map from previous step.

Figure 2 shows an overview of our method. Concretely, given a text prompt t that describes the action of a person, we first encode it to text embeddings z_t, and then use a text-to-image Stable Diffusion model ($\mathcal{G}$) to generate an image $\mathcal{I} = \mathcal{G}(z_t)$. The image is passed to encoder $\mathcal{E}_{\mathcal{G}_d}$ of the compression module of $\mathcal{G}_d$, a depth-to-image diffusion model, to get the compressed image latents $z \in \mathbb{R}^{4\times 64\times 64}$, i.e.

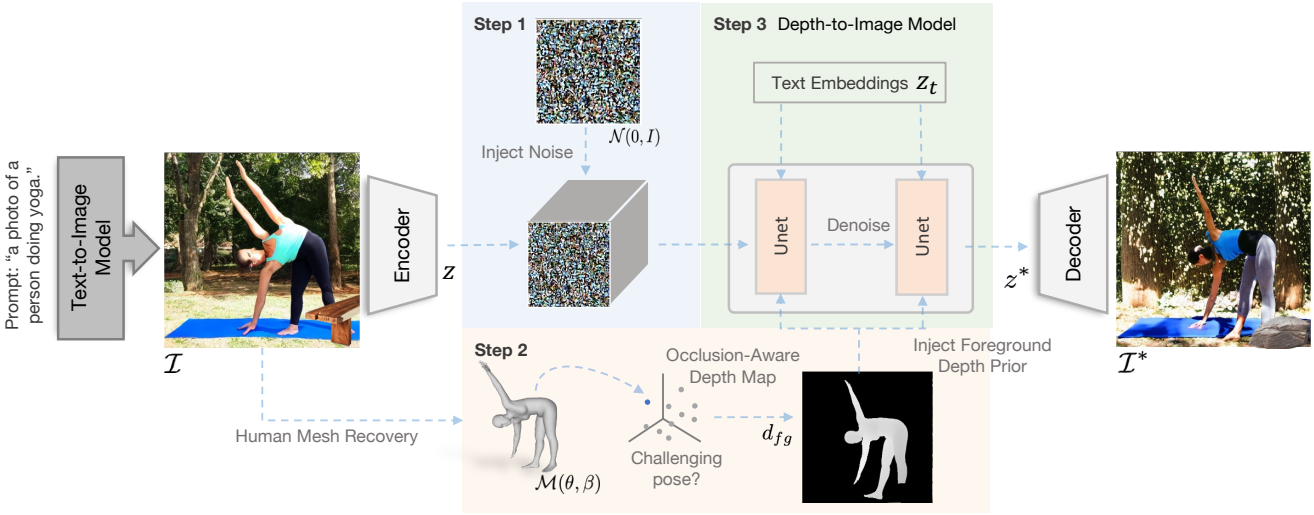

Figure 2. Overview of Diffusion-HPC. The generation process can be broken down into 3 steps. Step 1: Obtaining image latents z from the initial generation $\mathcal{I}$ of a pre-trained text-to-image model (i.e. Stable Diffusion [14]) and injecting noise. Step 2: Estimating human body mesh $\mathcal{M}(\theta, \beta)$ from $\mathcal{I}$. If the pose is challenging based on a pose prior (i.e. VPoser [38]) then render the mesh's depth map d_{fg} and introduce occlusions via object masks obtained from a segmentation model. Step 3: Using the latents z, foreground depths, and the text embeddings t as guide for the final generation $\mathcal{I}^*$.

$z = \mathcal{E}_{\mathcal{G}_d}(\mathcal{I})$. Image latents z contain context information about $\mathcal{I}$ such as the texture of the image and background layout. We can use z as a starting point in the final generation process so the context from the inital generation is roughly preserved.

Next, we use an off-the-shelf model to reconstruct the 3D mesh of the person in $\mathcal{I}$. Specifically, we estimate the human pose θ, shape β, and parameters of a weak perspective camera Π from the image using an off-the-shelf HMR model $f : \mathcal{I} \rightarrow (\theta, \beta, \Pi)$. Since our method focuses on rectifying implausible human generations, we determine whether the initial generation is likely to contain implausible humans and only apply our rectification process on images with hard poses. We observe that Stable Diffusion tend to generate worse anatomy on more difficult poses. Hence, we use a pre-trained human pose prior VPoser [38] as a proxy for determining if the person in image $\mathcal{I}$ has a challenging pose. VPoser is a Variational Auto-Encoder (VAE) that is trained on a massive database of realistic human poses [33]. By design, poses that are farther away from the canonical pose (i.e. challenging poses) have larger variance in the embedding space. Therefore, we identify a difficult pose θ if its embedding e_θ have larger norm, i.e. $||e_\theta||_2 > \tau$, where $\{\mu, \sigma\} = \mathcal{E}_v(\theta)$ and $e_\theta \sim \mathcal{N}(\mu, \sigma)$. $\mathcal{E}_v$ is the encoder of VPoser. τ is determined empirically and set to 30.

Now that we have the predicted human pose from $\mathcal{I}$, we move on to the final step of our method where we inject pose information $\mathcal{M}(\theta, \beta)$ into the generation process to produce a more plausible image of a person with the predicted pose θ. We achieve this by leveraging a depth-to-image version of Stable Diffusion, and using the depth values of the predicted human body as conditioning information in the generation process. Specifically, we render the 3D mesh to obtain the depth map $d_{fg} \in \mathbb{R}^{64\times64}$. Since there might be other objects in the image that occlude part of the person, we use Mask R-CNN [17] (pre-trained on COCO [31]) to segment the non-human objects in the image and use the segmentation masks to mask out the occluded body part in the depth map d_{fg}. Formally,

$$\{\theta, \beta, \Pi\} = f(\mathcal{I}) \tag{1}$$

$$d_{fg} = \mathcal{R}_d(\Pi, \mathcal{M}(\theta, \beta)) \tag{2}$$

$$d^*_{fg} = d_{fg} \odot ((1 - m) \cap (d_{fg} > 0)) \tag{3}$$

where $\mathcal{R}_d$ is a depth renderer that renders the depth map of a mesh, $\odot$ denotes the Hadamard product, and $(d_{fg} > 0)$ is the silhouette of the rendered person.

Finally, to preserve the context information (e.g. texture and background layout) of the initial generation, we use initial image latents as a starting point in the final generation process. We add noise to z, and use a pre-trained denoising model (i.e. depth-to-image Stable Diffusion) to perform sequential denoising steps which produces the final image latents z^*. The denoising process (achieved through a pre-trained UNet [42]) is guided by both the depth map d^*_{fg} and text embeddings z_t. Final generation is obtained by decod-

ing the z^* with the compression module's decoder $\mathcal{D}_{\mathcal{G}_d}$.

$$z^{noised} = noise(z) \tag{4}$$

$$z^* = denoise(z^{noised}; d^*_{fg}, z_t) \tag{5}$$

$$\mathcal{I}^* = \mathcal{D}_{\mathcal{G}_d}(z^*) \tag{6}$$

As shown in Figure 2, final generation $\mathcal{I}^*$ contains similar texture and background as the original image, but the human body anatomy is rectified.

3.3. Finetuning Human Mesh Recovery on challenging domains using synthetic data

Training a single-view Human Mesh Recovery (HMR) model end-to-end would require large amounts of images with paired 3D ground truths. Collecting such training sets requires burdensome motion capturing systems and is often limited to indoor laboratories. As a result, previous works such as [51] have focused on finetuning HMR model to a particular challenging domain using weak supervision (image paired with 2D keypoints). In this section, we introduce how image-mesh pairs from Diffusion-HPC can be used to finetuning HMR models in challenging domains.

Given a pre-trained HMR model that predicts pose θ, shape β and camera matrix Π from an image $\mathcal{I}$ (i.e. $f : \mathcal{I} \rightarrow (\theta, \beta, \Pi)$), we aim to adapt the model to a new target-domain by finetuning f on a small set of target images.

In a typical finetuning setup where only 2D keypoints from the target are available as supervision, 2D reprojection loss can be minimized to encourage the consistency between predicted and ground truth keypoints. Formally, for an image from the target training set, let the ground truth 2D keypoints be $j \in \mathbb{R}^{k\times 2}$ with k annotated keypoints per person, we would want to minimize $\mathcal{L}^{real}_{2D} = ||\hat{j} - j||_2$ where $\hat{j} = \Pi(\mathcal{W}\mathcal{M}(\hat{\theta}, \hat{\beta}))$ are the predicted 2D keypoints. Recall that $\mathcal{W}$ is the SMPL joint regressor, and Π is the projection matrix of a weak perspective camera.

Given this task setting, Diffusion-HPC can be used to generate synthetic data that has image-mesh pairs ($\mathcal{I}^*$ and $\{\theta, \beta, \Pi\}$ in Equations 1 and 6). Then, on those synthetic image-mesh pairs, we can supervise the model with ground truth body parameters, which provide stronger form of supervision as compared to 2D keypoints.

$$\mathcal{L}^{syn}_{3D} = ||\hat{\beta} - \beta||_2 + ||\hat{\theta} - \theta||_2 \tag{7}$$

Overall, loss during finetuning is $\mathcal{L} = \mathcal{L}^{real}_{2D} + \mathcal{L}^{syn}_{3D}$.

Guidance from real images. In the case of a clear target data domain for the HMR task, it can be useful to produce training data that have similar appearances to the target training set. Specifically, instead of using T2I diffusion model to generate the initial $\mathcal{I}$, we use real images from the training set. This guidance helps reduce the domain gap between the generated and real images because the appearances and poses will be more similar to the expected ones.

Pose augmentation. Finally, we can further enhance the diversity of the generated poses, by applying pose augmentations to the predicted poses. Specifically, after Equation 1, we can augment θ before proceeding to Equation 2. Formally, we apply pose augmentation in the embedding space of VPoser as in Weng et al. [51],

$$\mu, \sigma = \mathcal{E}_v(\theta) \tag{8}$$

$$z^{aug}_\theta = z_\theta \odot (1 + s\epsilon), z_\theta \sim \mathcal{N}(\mu, \sigma) \tag{9}$$

$$\theta^{aug} = \mathcal{D}_v(z^{aug}_\theta) \tag{10}$$

where $\mathcal{E}_v$ and $\mathcal{D}_v$ are the encoder and decoder of VPoser, s is a constant scalar, and ϵ is from a multivariate uniform distribution of the same dimension as the VPoser latent space.

4. Experiments

We first show the effectiveness of Diffusion-HPC for improving HMR performance in challenging domains in Sec. 4.1. Then, in Sec. 4.2 we present comparisons on the synthetic data generation quality of Diffusion-HPC.

4.1. Finetuning on challenging HMR settings

We demonstrate the potential of Diffusion-HPC through the task of few-shot adaptation of human mesh recovery models. We consider the setting where a small set of real images with 2D keypoints are available. This represents a typical scenario where we want to deploy a pre-trained HMR model on a new domain but there is limited ground truth annotations on the target domain. Through our experiments, we show that training with synthetic data from Diffusion-HPC improves HMR on challenging target domains as compared to previous adaptation methods.

We use the following sports datasets as they contain much more challenging poses than common HMR benchmarks. As a result, there is a large domain gap when applying pre-trained HMR models on those datasets, and finetuning is necessarily to close the domain gap. Pre-processing details are in the *Supplementary Material*.

- **Ski-Pose [41, 46]** includes 3D and 2D keypoints labels from 5 professional ski athletes in motion. There is a significant domain gap between ski poses and poses from other human pose estimation datasets, therefore Ski-Pose has been used as a benchmark in evaluating pose domain adaptation [15].
- **Sports Motion and Recognition Tasks (SMART) [9]** contains videos with per-frame 2D keypoints for various competitive sports. We consider 6 publicly released categories except for "badminton", which only contains one clip. We sample enough clips so that the training set contains roughly 100 images per category, and evaluate our finetuned models on the remaining images.

Evaluation metrics. For Ski-Pose, we use Mean Per Joint Position Error (MPJPE) and Procrustes-Aligned MPJPE

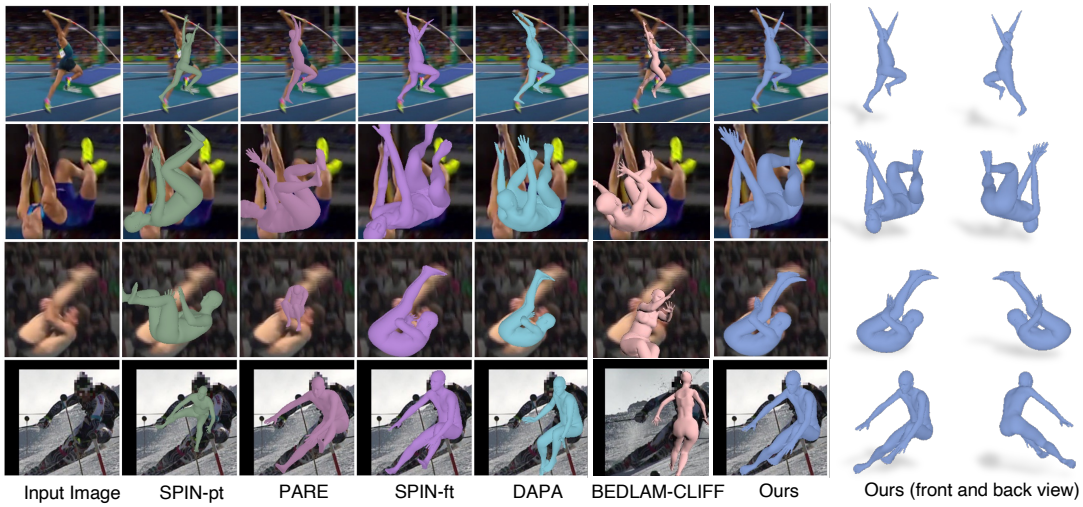

Figure 3. Qualitative HMR results on SMART and Ski-Pose datasets. Finetuning with data from Diffusion-HPC (rightmost) helps HMR models learn novel poses from challenging domains.

(PA-MPJPE) as our evaluation metrics. PA-MPJPE measures MPJPE after performing Procrustes alignment of the predicted and ground truth keypoints. SMART does not have ground truth 3D keypoints, so we report Percentage of Correct Keypoint (PCK) determined by distance between predicted and ground truth keypoints in pixels.

Implementation details. We use the backbone of SPIN [29] to estimate the human mesh, since the backbone is shared by both SPIN and DAPA [51], which enables fair comparison to these two. For each real image in the few-shot training set, we create 3 synthetic images, where each one has a slightly different pose due to pose augmentation. We finetune and update the entire HMR model with batch size of 64, learning rate of 1e-4. All hyperparameters are the same as in SPIN. The models are trained until the loss curves plateau and on average each finetuning experiment takes about 6 hours on a single NVIDIA TITAN V GPU.

Results. We compare to recent HMR models BEV [47] and PARE [28] that are pre-trained on MoCap datasets [24] as well as in-the-wild pose estimation datasets [2, 25, 31]. We also compare to BEDLAM-CLIFF [6], a state-of-the-art HMR model that is trained with a large synthetic dataset BEDLAM with realistic humans. In addition, we compare to finetuning methods SPIN-ft [29], DAPA [51], as well as finetuning with synthetic data generated with ControlNet [56] and Diffusion-HPC. The finetuning of these methods and ours minimizes 2D keypoint reprojection error by using 2D keypoints from the target training set. In addition, SPIN-ft uses in-the-loop model fitting to provide additional model-based supervision. DAPA generates synthetic data with paired 3D ground truths on the fly as additional supervision, while Ours uses data from Diffusion-HPC.

In Table 2 we report PCK on sports categories from SMART. Although the off-the-shelf models (SPIN-pt, BEV, PARE) were pre-trained on 2D datasets that include sports poses [25], there is still a significant domain gap between the training sets and SMART. BEDLAM-CLIFF was trained with a massive synthetic dataset BEDLAM, but the data generation was not tailored for the specific target domains, and therefore their training does not improve the model performance on the target dataset. As shown in the lower half of Table 2, finetuning on a small set of target images is helpful in closing the domain gap. Among those methods, we achieve better performance in general.

In Table 1, we report MPJPE/PA-MPJPE on Ski-Pose testset. We vary the size of the real training set during adaptation, and observe that with the same mount of real data, models trained with our synthetic data attain best performance. Further, with the help of synthetic data generated by Diffusion-HPC, we attain better performance than SPIN-ft and DAPA using much smaller amount of real data. We achieve best performance when using the entire training set. Notably, our best performance (111.3 MPJPE, 81.5 PA-MPJPE) is better than ProHMR [30] (122.7 MPJPE, 82.6 PA-MPJPE), which uses ground truth 2D keypoints from the testset as additional information. Finally, as qualitatively demonstrated in Figure 3, our method produces more accurate human mesh estimations on challenging poses, and in general have better alignment with 2D images.

4.2. Image generation quality

Data generation details. We use a text-to-image Stable Diffusion [14] model pre-trained on LAION-5B [44] and a CLIP ViT-L/14 [39] as text encoder. To condition the generation on depth maps we employ the depth-to-image Stable Diffusion model that was resumed from the text-to-image

Method	Ft.	Ft. with *syn*	PCK (↑) per Action						Mean
			Diving	Pole Vault	High Jump	Uneven Bars	Balance Beam	Vault	
SPIN [41]	✗	-	63.1	60.0	73.3	36.2	74.2	61.4	50.4
BEV [47]	✗	-	55.9	52.5	68.8	12.9	62.0	38.9	48.5
PARE [28]	✗	-	63.3	65.2	77.9	31.8	71.2	53.9	60.5
BEDLAM-CLIFF [6]	✗	-	30.4	57.3	67.4	31.1	55.7	48.3	48.4
SPIN-ft [41]	✓	✗	74.3	73.5	78.1	41.9	84.1	64.3	69.3
DAPA [51]	✓	✓	70.9	64.8	**79.4**	42.0	79.4	64.5	66.8
ControlNet [56]	✓	✓	70.6	65.3	74.2	43.4	83.6	62.3	66.6
Ours	✓	✓	**79.2**	**77.7**	78.1	**44.1**	**85.1**	**66.9**	**71.9**

Table 1. Quantitative results (PCK) on SMART. Ft indicates fine tuning on test data. Best numbers are in **bold**.

Method	Ft.	Ft. with *syn*	MPJPE (↓) / PA-MPJPE (↓)				
			0% train	1% train	5% train	50% train	100% train
SPIN [29]	✗	-	225.1 / 120.2	-	-	-	-
BEV [47]	✗	-	313.5 / 125.1	-	-	-	-
PARE [28]	✗	-	234.9 / 113.6	-	-	-	-
ProHMR [30]	✗	-	**122.7 / 82.6***	-	-	-	-
BEDLAM-CLIFF [6]	✗	-	363.5 / 136.5	-	-	-	-
SPIN-ft [41]	✓	✗	-	206.9 / 115.6	161.3 / 103.6	127.8 / 91.7	133.7 / 92.3
DAPA [51]	✓	✓	-	222.2 / 123.6	180.2 / 108.7	128.7 / 90.5	126.9 / 86.1
ControlNet [56]	✓	✓	-	194.6 / 106.2	144.1 / 94.0	118.2 / 85.3	114.2 / 83.4
Ours	✓	✓	-	**182.3 / 105.9**	**143.4 / 90.7**	**116.5 / 83.5**	**111.3 / 81.5**

Table 2. Quantitative comparisons on Ski-Pose. We report MPJPE/PA-MPJPE on the test set. (*: Note that ProHMR uses ground truth 2D keypoints for test-time optimization and therefore has an unfair advantage for this experiment. All other models (including Ours) perform direct inference on test set). The best number per configuration is in **bold**.

model, and finetuned for 200k steps. The denoising model has an extra input channel to process the (relative) depth prediction produced by MiDaS [40] which is used as added conditioning. As our segmentation model we use Mask R-CNN [17, 53] pre-trained on MS-COCO [31]. For the qualitative examples in Figure 4 and experiments in Section 4.2, we use BEV [47] as the HMR model, due to its capacity of recovering people of all age groups and better empirical performance at localizing implausible synthetic humans, whereas two-stage HMR models that rely on a human detector often treat these erroneous generations as false negatives. With 50 inference steps, it takes about 6 seconds to create an image starting from text, and in the setting when a real image is used as guidance, the time is halved.

4.2.1 Comparison on text-conditioned generation

We assess the quality of the text-only conditioned images generated by Diffusion-HPC by comparing them to off-the-shelf Stable Diffusion. In order to span a wide taxonomy of human activities we compose text prompts from the category labels available in the MPII [2] dataset. In addition, to assess the generation quality regarding extremely challenging human poses, we use the publicly released sports categories from SMART [9] (further introduced in Sec. 4.1) as text prompts. For both datasets, we report the standard evaluation metric Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) [20]. Since the focus of our method is on human generation, we report H-FID / H-KID, which is FID / KID computed with only foreground humans (segmented by Mask R-CNN). Note that FID/KID are computed using image-level features, and therefore do not focus on human generation quality in particular. Thus, we deem H-FID/H-KID more suitable metrics for our work.

Model	Dataset	User Preference (↑)	FID / H-FID (↓)	KID / H-KID (↓)
Stable Diffusion	MPII	0.45 ± 0.23	**75.6** / 70.5	**0.03** / 0.11
Diffusion-HPC	MPII	**0.55 ± 0.23**	**75.6 / 68.1**	**0.03 / 0.04**
Stable Diffusion	SMART	0.23 ± 0.08	**66.3** / 91.4	0.04 / 0.07
Diffusion-HPC	SMART	**0.77 ± 0.09**	67.7 / **89.5**	**0.03 / 0.06**

Table 3. Text-conditioned comparisons on activities from MPII.

Furthermore, we perform a user study where 6 independent blinded users were shown a randomly sampled set of 100 side-by-side images each generated by Stable Diffusion and Diffusion-HPC. The users were given the task of selecting the image with the most plausible human pose and anatomy. If the images were comparable, the user could select a "no preference" option.

Results. Table 3 presents comparisons on text-conditioned generations. While FID/KID values are roughly the same, we highlight that humans generated by Diffusion-HPC have lower H-FID/H-KID to humans from real images. User study suggests that users prefer our gen-

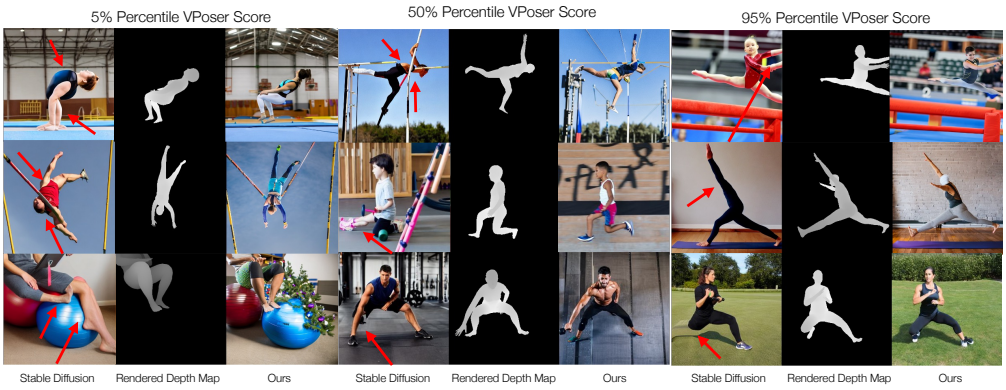

Figure 4. Comparison with Stable Diffusion [14] on text-conditioned generations. Red arrows point out implausible body parts in Stable Diffusion generations. To show a spectrum of varying pose difficulty levels, we present generations from the 5%, 50%, 95% quantiles (i.e. from easy to hard) in terms of VPoser score. Rendered depths are included to show correct pose guidance.

erations most of the time. Qualitative results in Figure 3 suggest that our generations, while preserving the textures of the original images (hence similar FID/KID), effectively corrects the human anatomy (hence lower H-FID/H-KID).

4.2.2 Comparison on pose-conditioned generation

Most previous pose-conditioned generative models focus on the task of "reposing" [1, 27, 34], where the goal is to repose the reference person using the target pose. These models are trained on fashion catalog images with clean background, therefore they are too simplistic to be effective baselines for our purpose. The only fair baseline, to our knowledge, is Brooks and Efros [7], a recent StyleGAN2 [26]-based generative model that takes 2D keypoints of a posed person and generates images with compatible background. We benchmark their pre-trained model (trained on 18 million images sourced from 10 existing human pose estimation and action recognition datasets) on MPII for in-domain assessment.

Results. Table 4 shows quantitative comparisons of image quality. Notably, even though Brooks and Efros [7] was trained with paired data (keypoint-image pairs) and therefore has an advantage over Diffusion-HPC where the underlying models are trained/finetuned only with images, Diffusion-HPC consistently achieves better performance. Moreover, Brooks and Efros [7] has poor generalization capability to novel pose distributions as in SMART, whereas our method powered by Stable Diffusion has a better zero-shot capability. Additional details, qualitative comparisons and limitations are in the *Supplementary Material*.

Method	Trained with paired data	T	R	D	FID/ H-FID	KID/ H-KID
Results on MPII						
Brooks et al.	✓	✗	✗	✓	109.0 / 75.5	0.10 / 0.07
Diffusion-HPC	✗	✗	✓	✗	95.6 / 59.3	0.07 / 0.05
	✗	✓	✗	✗	54.6 / 41.3	0.03 / 0.03
	✗	✓	✓	✗	44.6 / 41.5	0.02 / 0.03
	✗	✗	✓	✓	95.3 / 58.1	0.07 / 0.04
	✗	✓	✗	✓	72.8 / 136.2	0.03 / 0.11
	✗	✓	✓	✓	**42.6 / 38.6**	**0.02 / 0.02**
Results on SMART						
Brooks et al.	✓	✗	✗	✗	175.8 / 114.1	0.14 / 0.06
Diffusion-HPC	✗	✗	✓	✗	85.9 / 121.5	0.06 / 0.07
	✗	✓	✗	✗	162.3 / 113.9	0.13 / 0.08
	✗	✓	✓	✗	94.8 / 99.8	0.06 / 0.05
	✗	✗	✓	✓	**85.5** / 122.2	0.06 / 0.07
	✗	✓	✗	✓	145.9 / 131.5	0.10 / 0.07
	✗	✓	✓	✓	92.4 / **44.5**	**0.06 / 0.03**

Table 4. Pose-conditioned generation quality. Note that Brooks and Efros [7] was trained with paired data (images with corresponding 2D keypoints), whereas Diffusion-HPC is trained/finetuned only with images. Diffusion-HPC can take the background of text ("T") and/or a real image ("R") as conditioning information.

5. Conclusion

We proposed Diffusion-HPC, a text-conditioned and training-free method that injects model-based human body prior to improve human-centric generation of state-of-the-art text-conditioned and pose-conditioned generative models. Further, Diffusion-HPC demonstrates excellent utility in a challenging downstream task, single-view HMR. For future work, we anticipate further investigation into the obstacles associated with human generation in foundation generative models, as well as exploring innovative ways of using generative models to tackle the challenges in 3D human perception tasks.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. 2026498. L. Bravo-Sánchez acknowledges financial support for this work by the Fulbright U.S. Student Program, which is sponsored by the U.S. Department of State and Fulbright Colombia. In addition, she was partially supported by an educational grant from IBM Research.

Diffusion-HPC: Synthetic Data Generation for Human Mesh Recovery in Challenging Domains

Supplementary Material

A. Code

Code and trained models can be found at https://github.com/ZZWENG/Diffusion_HPC.

B. Additional Implementation Details

Quantitative evaluation in Table 1. We compute quantitative metrics using roughly 10,000 images generated from MPII [2] and SMART [9] prompts, respectively.

For MPII, we use "{image description} of {person} doing {action}" as the text prompts, where "{person}" can be single- or multi-person descriptions of person(s) of interest, and "{action}" are the activity categories from MPII. (We exclude categories "inactivity, quite/light" and miscellaneous" because they do not describe a specific activity.) For example, resulting prompts could be "a nice photo of a man doing water activities." or "a high-resolution photo of a group of people doing conditioning exercises".

Text prompts for SMART are constructed using the template "a photo of an athlete doing {action}" where action is one of "high jump", "vault", "pole vault", "diving", "gymnastics on uneven bars", and "gymnastics on a balance beam".

Data processing for downstream experiments. Following previous works [29, 51], we crop the images such that the persons (localized by ground truth 2D keypoints) are centered in the crop. In addition, the persons are scaled such that the torso (i.e. mean distance between left/right shoulder and hip) are roughly one third of the crop size (224×224).

C. Additional Comparisons on Pose-Conditioned Generation

Effect of text and real guidance. Figure S5 demonstrates qualitative comparisons between different versions of our model and Brooks and Efros [7] As shown, text guidance (T) is essential in capturing the context of the human action. Guidance from real images (R) provides overall texture information such as background colors. While guidance from real images alone is not sufficient in preserving the action of the human (3_{rd} row in Ours R), it adds to text guidance, and further improves the realism of the image generations (Ours T+R).

Effect of finetuning. To see whether a finetuned diffusion model could further help improve generation quality, we finetune Stable Diffusion on the target dataset (MPII and SMART respectively) for 10 epochs. Generations with finetuned diffusion models is noted with "D". As shown in Figure S5, although finetuned diffusion model generates images with better background when there is no real guidance (Ours T vs. T+D), the foreground often loses the texture of humans, which is likely due to the "catastrophic forgetting" as sometimes observed in finetuning large pretrained models. Qualitatively, with both text and real guidance, the effect of finetuning is barely noticeable (Ours T+R vs. T+R+D). Quantitatively, when using real guidance (with or without text guidance), finetuning slightly improves FID, and significantly improves H-FID and H-KID. Further, consistent to what is observed in qualitative results, 41.3 H-FID (with T) vs. 136.2 H-FID (with T+D) suggests that finetuning worsens performance without real guidance. This suggests that for text-conditioned generations, it is optimal to utilize an off-the-shelf diffusion model without finetuning.

D. Additional Qualitative Results

Here we include additional qualitative results as well as failure cases for the text-conditioned and pose-conditioned generations (Section 4.2).

Text-Conditioned Generation

Figure S6 shows qualitative comparisons of Stable Diffusion [14] and Diffusion-HPC on text-conditioned generations. The images were selected from those sampled for the user study.

In Figure S7 we include typical failure cases of text-conditioned generations. In left and middle columns, the body structures are not sufficiently rectified. This is likely because that resolution of depth maps (used for conditioning) is limited (64×64), so consequently small humans with out-of-distribution poses are challenging to rectify. In the right column, we show a failure scenario when the HMR model (i.e. BEV [47]) fails to reconstruct the humans in close-up shots. We could consider filtering out close-up shots, as they are not the primary intended use cases for Diffusion-HPC.

Pose-Conditioned Generation

Figure S9 shows qualitative comparisons of Brooks and Efros [7] and Diffusion-HPC on pose-conditioned generations, and Figure S10 shows failures cases of pose-conditione generations. As seen from "Ours T+R" and "Ours T+R+D", human-object interactions are sometimes not preserved.

Figure S5. Qualitative comparisons to Brooks and Efros [7] (input 2D keypoints are overlaid on the bottom left). Our generations conditioned on text (T), real images (R), and in-domain (D).

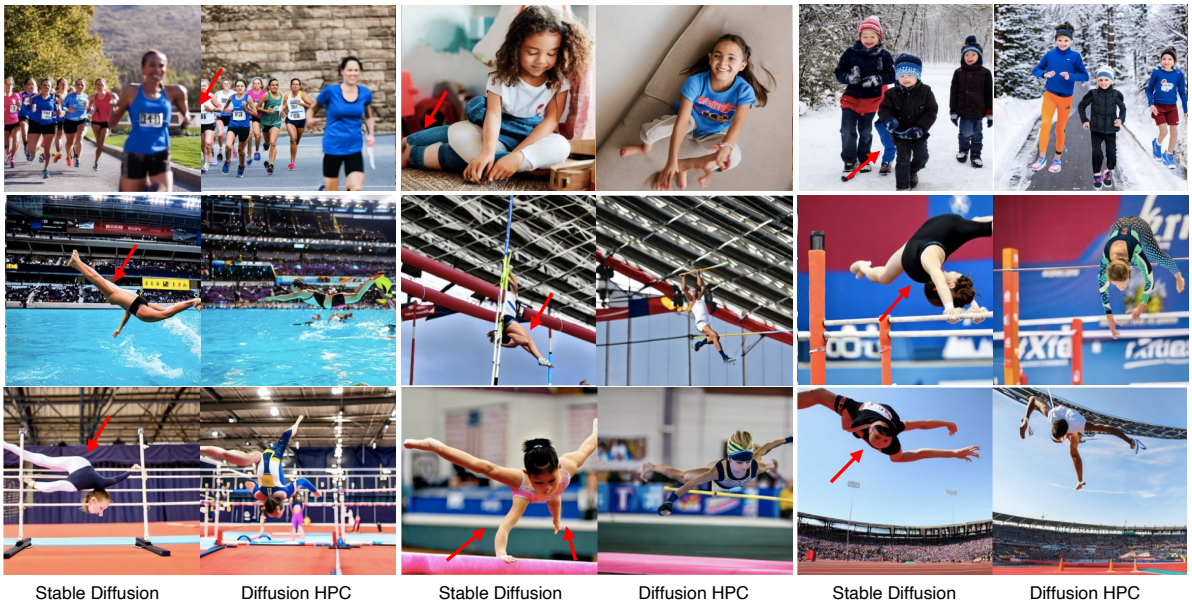

Figure S6. Comparison with Stable Diffusion [14] on text-conditioned generations. Row 1 and rows 2-3 are generated with MPII [2] and SMART [9] prompts, respectively. Red arrows point out implausible body parts in Stable Diffusion generations.

Note that in Diffusion-HPC, human-object interactions are considered but not modelled in an explicit way. Specifically, when we construct the depth map, we use Mask R-CNN [17] to segment out the occluded body part, which helps with scenarios when, for instance, the person is riding the horse (row 1 of Figure S9). However, row 2 of figure S10 shows a failure case where the boat is not detected by Mask R-CNN.

In addition, in Diffusion-HPC, latents from the initial generations help preserve the objects and context in the fi-

Figure S7. Failure cases on text-conditioned generations.

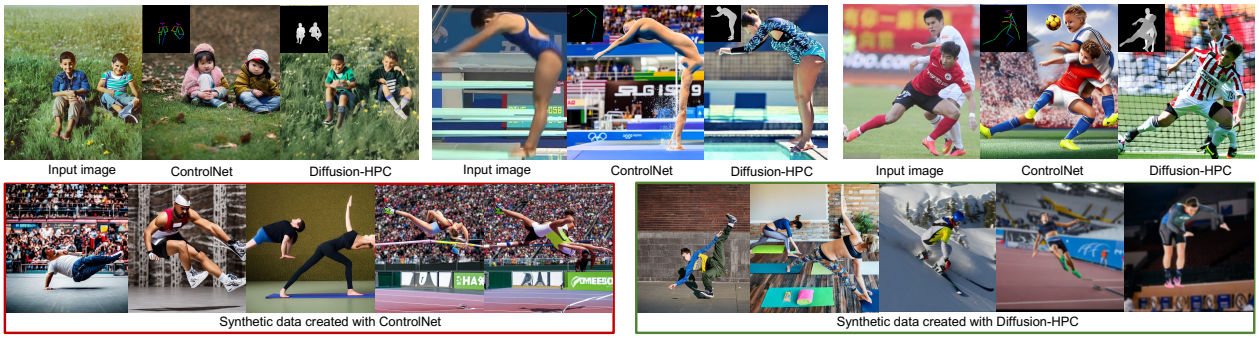

Figure S8. Comparison to images generated by ControlNet. Top: Pose-to-image ControlNet often results in misaligned limbs (e.g. child legs in top left; front person's legs in top right), and does not generalize to challenging domains such as sports poses in general. Bottom: Example synthetic images generated by pose-conditioned ControlNet and Diffusion-HPC. We highlight Diffusion-HPC's superior generalization capability, even on extremely challenging domains such as sports. Such generalization capability is key for it to be effective for downstream tasks, such as HMR.

nal generated scenes. For the pose-conditioned generations here, latents of real images help capture the background objects such as horse and surfboard (row 1 and 3 of Figure S9). However, when the background object is occluded or small (row 2 in Figure S9 and row 1 in Figure S10), the latents are not sufficient in preserving the object in the final generations. Future work could consider extending Diffusion-HPC by explicitly modelling the human-object/scene interaction.

Comparison to ControlNet [56]. Adding control to pre-trained generative models has received increasing attention. While it is possible to use works such as ControlNet [56] to generate pose-conditioned images, we note that ControlNet needs additional finetuning, requiring greater computing and annotation resources. In particular, training pose-to-image ControlNet requires paired data (i.e. 2D keypoints and images) which also limits the training data distribution to easy poses. 2D keypoints inherently provide less information compared to 3D body pose, and solely relying on them for 3D pose understanding tasks is insufficient. As a comparison, we train HMR models on SMART and SkiPose using the synthetic data generated with ControlNet, and on both evaluation sets, PCK and PA-MPJPE are worse than training with Diffusion-HPC (Table 1 and 2). Notably, for SMART, training with ControlNet data is worse than SPIN-ft which does not use synthetic data at all. Regardless of the possibility of re-training ControlNet using 3D human representations as conditioning, it is impractical when considering the substantial number of images with paired 3D GTs demanded. For reference, ControlNet pose-to-image model was trained on 200K keypoint-image pairs. Note that the scarcity of such 3D data was the primary motivation behind our work. Thus, rather than perceiving ControlNet as a direct alternative to our approach, it is more fair to consider it as a promising avenue to enhance our work, as the two methods can be synergistically combined - we could use Diffusion-HPC to bootstrap large amounts of image data with paired 3D pseudo ground truths and then use ControlNet to finetune a diffusion model.

E. Limitations

As we rely on large pre-trained models [14, 44], any biases in these models or datasets that they were trained on will be replicated onto our generated images. Due to the resolution of depth maps (64×64), fine details such as fingers and facial expressions are challenging to synthesize. Besides, since we only render person depth maps, human-object/human-scene interactions may not be well-preserved in the final generation (e.g. the person and yoga mat in column 3, row 2 of Figure 3). While these limitations do not affect downstream tasks where we only care about the body pose, there is large room to improve the photo-realism of

Figure S9. Additional qualitative comparisons to Brooks and Efros [7] on the MPII dataset. Input 2D keypoints to Brooks and Efros [7] are overlayed on the bottom left in column 2. Top 3 rows are from MPII, and bottom 3 rows are from SMART. Our generations conditioned on text (T), real images (R). "(D)" means the diffusion model is finetuned on the target dataset (MPII and SMART respectively).

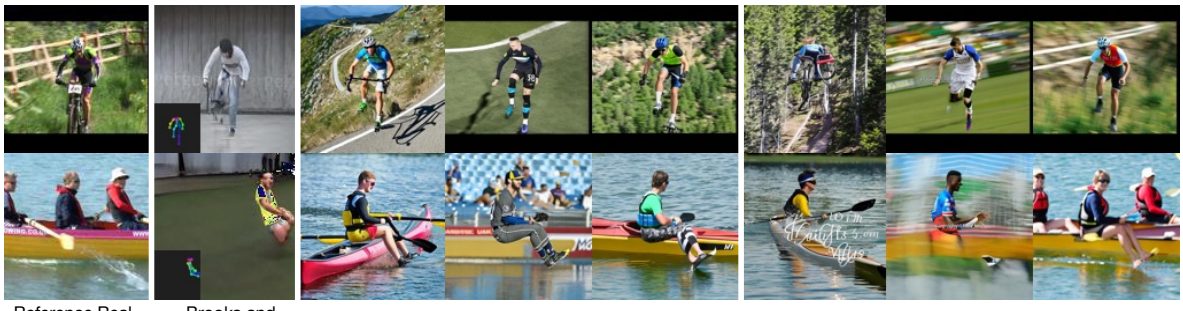

Figure S10. Failure cases on pose-conditioned generations.

human-centric image synthesis, and for the synthetic data to be useful for a wider variety of downstream tasks such as expressive HMR [38] and recovering human-object/scene interaction [5, 50, 55]. Lastly, as we use SMPL body representation, our method does not consider people with limb losses, but it can be adapted to do so.

References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021. 3, 8

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 6, 7, 1, 2

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 3

[4] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 3

[5] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 4

[6] 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)*, pages 8726–8737, 2023. 2, 6, 7

[7] Tim Brooks and Alexei A Efros. Hallucinating pose-compatible scenes. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 510–528. Springer, 2022. 3, 8, 1, 2, 4

[8] 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. *arXiv preprint arXiv:2304.00916*, 2023. 2

[9] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129:2846–2864, 2021. 2, 5, 7, 1

[10] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 346–362. Springer, 2022. 2, 3

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 3

[12] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[13] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. *arXiv preprint arXiv:2305.02312*, 2023. 2

[14] Rombach *et al*. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4, 6, 8

[15] Mohsen Gholami, Bastian Wandt, Helge Rhodin, Rabab Ward, and Z Jane Wang. Adaptpose: Cross-dataset adaptation for 3d human pose estimation by learnable motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2022. 5

[16] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2, 3

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4, 7, 2

[18] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 1, 2

[19] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 3

[22] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 2

[23] HuggingFace, 2022. 2, 3

[24] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 6

[25] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, page 5. Aberystwyth, UK, 2010. 6

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 8

[27] Markus Knoche, István Sárándi, and Bastian Leibe. Reposing humans by warping 3d features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1044–1045, 2020. 3, 8

[28] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 6, 7

[29] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 2, 6, 7, 1

[30] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11605–11614, 2021. 6, 7

[31] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4, 6, 7

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3

[33] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 3, 4

[34] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020. 3, 8

[35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[36] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 3

[37] 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 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 2

[38] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[40] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 7

[41] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8437–8446, 2018. 2, 5, 7

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[43] 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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 3

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 1, 2, 6, 3

[45] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, 2020. 2

[46] Jörg Spörri. Reasearch dedicated to sports injury prevention-the'sequence of prevention'on the example of alpine ski racing. *Habilitation with Venia Docendi in Biomechanics*, 1(2): 7, 2016. 2, 5

[47] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022. 6, 7, 1

[48] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 2

[49] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 2

[50] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 334–343, 2021. 4

[51] Zhenzhen Weng, Kuan-Chieh Wang, Angjoo Kanazawa, and Serena Yeung. Domain adaptive 3d pose augmentation for in-the-wild human mesh recovery. *International Conference on 3D Vision (3DV)*, 2022. 2, 5, 6, 7, 1

[52] Zhenzhen Weng, Zeyu Wang, and Serena Yeung. Zeroavatar: Zero-shot 3d avatar generation from a single image. *arXiv preprint arXiv:2305.16411*, 2023. 2

[53] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 7

[54] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023. 2

[55] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020. 4

[56] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3, 6, 7

[57] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 2