PoseSyn: Synthesizing Diverse 3D Pose Data from In-the-Wild 2D Data Supplementary

A. Implementation Details

Implementation Details We selected three image-to-3D pose estimation models as our target pose estimators (TPEs): Hybrik [1], 3DCrowdNet [2], and 4DHumans [3]. All experiments were conducted utilizing each official code of 3DCrowdNet[1], Hybrik[2] and 4DHumans[3].

Specifically, for 3DCrowdNet, we downloaded a pre-trained model which was trained for 10 epochs on real datasets (*e.g.*, Human36M [4], MuCo [5], MSCOCO [6], and MPII [7] datasets) with a learning rate of 1×10^{-4}. We then applied our data synthesis framework to augment the MPII dataset, creating a 3D pose dataset with 27,291 samples. We fine-tuned the 3DCrowdNet on this synthesized dataset with a batch size of 64 and a learning rate of 1×10^{-5} for 10 epochs, utilizing both real data and synthesized data.

In the case of 4DHumans model, we downloaded a pre-trained model which was trained for 1M iterations on real datasets (*e.g.*, Human36M, MPI-INF-3DHP [8], AVA [9], AIC [10], INSTA [11], MSCOCO, and MPII datasets) with a learning rate of 1×10^{-4}. We then applied our data synthesis framework to augment the MPII dataset, creating a 3D pose dataset with 27,872 samples. We fine-tuned the 4DHumans on this created dataset with a batch size of 32 and a learning rate of 1×10^{-5} for 200K iterations, utilizing both real data and created data.

Likewise for Hybrik, we downloaded a pre-trained model which was trained for 120 epochs on real datasets (*e.g.*, Human36M, MPI-INF-3DHP, and MSCOCO datasets) with a learning rate of 1×10^{-3}. We then applied our data synthesis framework to obtain a synthesized 3D pose dataset consisting of 26,758 samples, and fine-tuned the Hybrik with a batch size of 64 and a learning rate of 1×10^{-4} for 40 epochs utilizing both real data and synthesized data. Note that each synthesized 3D dataset differs per model because challenging images are extracted from each TPE.

Additional parameters, including the number of challenging data (*i.e.*, N_{C}), the number of non-challenging data (*i.e.*, N_{NC}), and the filtering threshold τ, are provided in Tab. 1.

VLM Prompting In our proposed Semantic-guided Motion Generation (SMG) stage, we augments an identified challenging pose into motion sequences guided by both textual description and initial pose representation. To extract the textual description of the challenging image, we leverage a VLM [12] to ask the question, *"What is the motion of the someone in the image? Please answer similar to {Text-to-Motion prompt examples}"*, where {*Text-to-Motion prompt examples*} are provided as follows:

– "A man kicks something or someone with his left leg."
– "A person walking forward and then turns around."
– "A person squats down then jumps."
– "A person raises their right hand to their face."
– "He is waving with his right hand."
– "A person kicks something with their right foot."
– "A running man hops over something a comes down to a walk."
– "A man raises both arms, then kneels down."
– "Person is using their left arm to dodge a punch."
– "A person raised both their arms and started to clap."

We visualize the results of VLM answers in Fig. 1.

[1] https://github.com/hongsukchoi/3DCrowdNet_RELEASE
[2] https://github.com/Jeff-sjtu/HybrIK
[3] https://github.com/shubham-goel/4D-Humans

Details for Orientation-aligned Motion Guidance PoseSyn synthesizes human animated images $\mathcal{V}_{\text{C}} = \{I_{\text{C},l}\}_{l=1}^{L}$ by using non-challenging image I_{NC} as reference image and generated motion sequences $\mathcal{M}_{\text{C}} = \{J_{\text{C},l}^{\text{3D}}\}_{l=1}^{L}$ as motion guidance. To ensure natural human animation, we align the global orientation of $\mathcal{M}_{\text{C}}$ with the human orientation in I_{NC}. First, following Champ [13]'s original method, we exploit 4DHumans to obtain camera parameters (*i.e.*, focal length f and principal points p) and SMPL parameters including global orientation $\theta^{\text{G}} \in \mathbb{R}^3$, pose $\theta^{\text{P}} \in \mathbb{R}^{69}$, shape $\beta \in \mathbb{R}^{10}$, and translation $\bar{t} \in \mathbb{R}^3$ for the reference image I_{NC}, which is denoted as:

$$\theta^{\text{G}}, \theta^{\text{P}}, \beta, \bar{t}, f, p = \text{4DHumans}(I_{\text{NC}}). \tag{1}$$

As Champ utilizes rendered meshes of the motion sequences as motion guidance, we obtain SMPL parameters $\{\theta_{\text{S},l}^{\text{G}}, \theta_{\text{S},l}^{\text{P}}\}_{l=1}^{L}$ from the 3D joint sequences $\mathcal{M}_{\text{C}}$ by applying an optimization method [14] O as follows:

$$\{\theta_{\text{S},l}^{\text{G}}, \theta_{\text{S},l}^{\text{P}}\}_{l=1}^{L} = O(\mathcal{M}_{\text{C}}), \tag{2}$$

where SMPL parameters are represented in the Rodrigues' rotation form [15]. To align the global orientation of $\mathcal{M}_{\text{C}}$ with the human orientation in I_{NC}, we use the global orientation θ^{G} of $\mathcal{M}_{\text{C}}$ as the initial global orientation for the first motion sequence. For the remaining frames, we adjust the initial global orientation based on the relative angle between the global orientations of the motion sequences. Specifically, let's denote $T_{r\to e}(\cdot)$ as the conversion from the Rodrigues' rotation form to Euclidean one, $R_{\text{S},l} = T_{r\to e}(\theta_{\text{S},l}^{\text{G}})$ as the rotation matrix for the global orientation represented in Euclidean space, and then the modified global orientation sequences are formulated as follows:

$$\theta_{\text{mod},l}^{\text{G}} = T_{r\to e}^{-1}(R_{\text{S},l}(R_{\text{S},l-1})^{-1}R_{\text{init}}), \quad l = 2, ..., L, \tag{3}$$

where $\theta_{\text{mod},1}^{\text{G}} = \theta^{\text{G}}$ is the initial global orientation and $R_{\text{init}} = T_{r\to e}(\theta^{\text{G}})$ is its representation in Euclidean space. Then, we acquire orientation-aligned mesh sequences $\{Mesh_l\}_{l=1}^{L}$, with each orientation-aligned mesh defined as:

$$Mesh_l = SMPL(\theta_{\text{mod},l}^{\text{G}}, \theta_{\text{S},l}^{\text{P}}, \beta, \bar{t}). \tag{4}$$

Finally, we project each mesh onto the image plane by using the camera parameters for the reference image I_{NC} to obtain a sequence of orientation-aligned rendered mesh $\{I_{\text{render},l}\}_{l=1}^{L}$, where each rendered mesh is calculated as follows:

$$I_{\text{render},l} = \text{Proj}(Mesh_l, f, p). \tag{5}$$

This rendered mesh sequence serves as motion guidance for animating a human in the reference image I_{NC}. For more details, please refer to [13].

Symbol	Description	dimension	Value
	Notation related to dataset		
$\mathcal{D}$	2D Pose Dataset. In our experiments, we used MPII dataset.	-	-
$\mathcal{D}_{\text{C}}$	A set of challenging dataset identified through EEM.	-	-
$\mathcal{D}_{\text{NC}}$	A set of non-challenging dataset identified through EEM.	-	-
	Notation related to image		
I	Input image	(256, 256, 3)	-
W	Image width	-	-
H	Image height	-	-
I_{C}	Challenging image	(256, 256, 3)	-
I_{NC}	Non-challenging image	(256, 256, 3)	-
	Notation related to joint		
N_{2D}	The number of 2D joints.	-	16
N_{3D}	The number of 3D joints.	-	24
$J_{\text{GT}}^{\text{2D}}$	Ground truth 2D pose.	(16, 2)	-
$\hat{J}^{\text{2D}}$	2D projection of the 3D pose predicted on the input image through TPE.	(16, 2)	-
$\hat{J}^{\text{3D}}$	3D pose predicted on the input image through TPE.	(24, 3)	-
$\hat{J}_{\text{C}}^{\text{3D}}$	3D pose predicted on the challenging image through TPE.	(24, 3)	-
$\hat{J}_{\text{NC}}^{\text{3D}}$	3D pose predicted on non-challenging image through TPE.	(24, 3)	-
$\hat{J}^{\text{2D},n}$	Body part index values of 2D pose (*e.g.*, right shoulder, right leg, left shoulder, etc.).	(2)	-
$\hat{J}^{\text{3D},n}$	Body part index values of 3D pose (*e.g.*, right shoulder, right leg, left shoulder, etc.).	(2)	-
	Notation related to EEM		
Err	Error value calculated by Eq. (1) to identify challenging dataset.	-	-
$\mathbf{w}_n$	Hyperparameter considering the importance of joint parts in Err value calculation.	(16)	Ankle(1), wrist(1), Elbow(0.5), knee(0.5), Hip(0.25), shoulder(0.25)
$\text{Top}_{K_{\text{C}}}$	Operation used to identify K_{C} challenging dataset $\mathcal{D}_{\text{C}}$.	-	500
$\text{Top}_{K_{\text{NC}}}$	Operation used to identify K_{NC} non-challenging dataset $\mathcal{D}_{\text{NC}}$.	-	200
	Notation related to PAM		
$\mathcal{MR}_{\text{init}}$	Initial motion representation obtained by repeating mis-predicted pose over time step T.	(30, 263)	-
$\mathcal{Z}_{\mathcal{MR}}$	Latent features of initial motion sequences obtained from encoder in VQ-VAE.	(7, 512)	-
$\mathcal{S}_{\mathcal{MR}}$	Index values of initial motion sequences obtained through codebook quantization.	(7)	-
F	Preprocessing method [16] for motion representation.	-	-
T	Time step for acquiring a initial motion representation from a initial pose.	-	30
$\mathcal{E}$	Encoder of Motion VQ-VAE in T2M-GPT [14].	-	-
M	Length of initial motion sequences obtained by dividing T by r.	-	7
r	Temporal downsampling factor of Encoder $\mathcal{E}$.	-	4
$\mathbf{e}_{\text{text}}$	Text embedding of textual description processed by CLIP encoder [17].	(512)	-
L	Length of augmented challenging motion sequence $\mathcal{M}_{\text{C}}$.	-	-
$\mathcal{M}_{\text{C}}$	Augmented challenging motion sequence obtained through transformer under condition of both $\mathbf{e}_{\text{text}}$ and $\mathcal{S}_{\mathcal{MR}}$.	$(L, 22)$	-
$\mathcal{V}_{\text{C}}$	Human animated video obtained through [13] with motion guidance $\mathcal{M}_{\text{C}}$ and reference image I_{NC}.	$(L, 256, 256, 3)$	-
	Notation related to Error & Filtering		
$Err_{3D,l}$	Error value calculated by Eq. (6) in filtering stage.	-	-
τ	Threshold value used in filtering stage.	-	120
	Remaining notations		
TPE	Top-down human pose estimation model (*e.g.*, 3DCrowdNet [2], Hybrik [1], and 4DHumans [3]).	-	-
f	Focal length used in 2D projection of 3D pose.	(2)	3DCrwodNet(5000), 4DHumans(5000 / 256 $\times$ W), Hybrik(1000 / 256 $\times$ W)
p	Principal point used in 2D projection of 3D pose.	(2)	Bounding box center

Table 1. **Notation Table.**

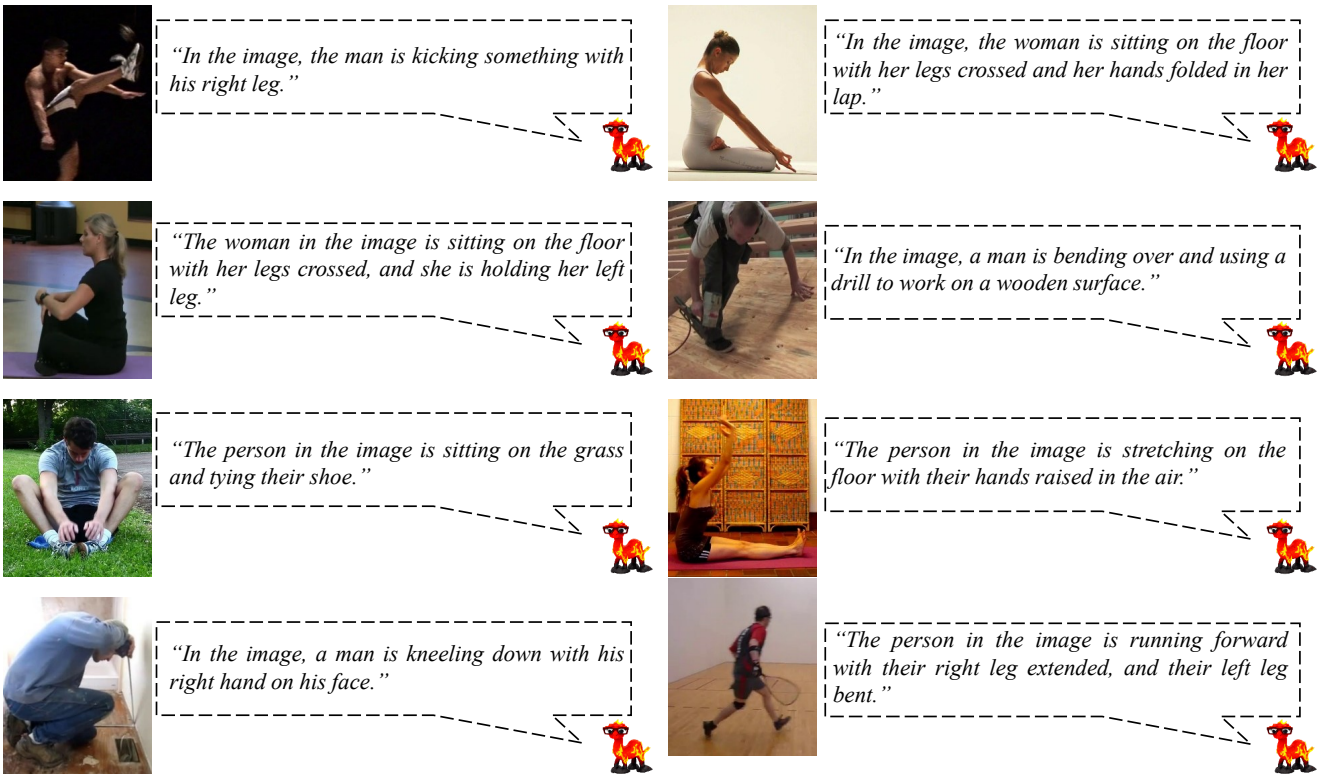


Figure 1. **VLM Prompting.** We visualize challenging image and extracted textual description pairs.

B. Extendibility of PoseSyn

To ensure a fair comparison with the baselines, we utilized the MPII dataset [7], which was originally used to pre-train the TPE (3DCrowdNet), as the basis for synthesizing the 3D pose dataset throughout our experiments. However, our methodology is not restricted to the data used for pre-training the TPE; rather, it can augment any easily accessible in-the-wild unlabeled images into a 3D pose dataset, showcasing its flexibility and broad applicability.

To validate this capability, we assigned pseudo-label annotations to each image in the dataset using either a 2D pose estimator [18] or a 3D pose estimator [19], rather than relying on the 2D GT poses from the MPII dataset. In the case of 2D pseudo-labeling, the EEM module calculated the error as outlined in Eq. (1), which was then used to identify challenging and non-challenging datasets as expressed in Eq. (2) and Eq. (3), respectively. For the 3D pseudo-labeling, the criterion metric Err was computed by replacing 2D joints with their 3D counterparts in Eq. (1), followed by the same methodology for identifying challenging and non-challenging samples.

	3DPW		EMDB		CMU_171204		CMU_171026		HuMMan		Mean	
	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
Real-only	81.7	51.1	115.8	71.2	108.8	72.5	110.7	70.4	98.9	65.8	103.2	66.2
3D Pose Estimator	78.9	49.8	112.5	69.9	103.6	69.5	106.6	69.9	96.5	63.9	99.6	64.6
2D Pose Estimator	78.5	49.7	112.2	69.9	103.3	69.6	106.4	68.5	**93.1**	63.7	98.7	64.3
Ground Truth	**77.4**	**48.9**	**111.0**	**68.3**	**101.0**	**67.3**	**105.0**	**67.9**	**93.1**	**62.3**	**97.5**	**62.9**

Table 2. **EEM Approach Variants.** We report results in the performance improvement of TPE when trained with synthesized dataset obtained through different EEM approaches: 3D pose pseudo-labeling, 2D pose pseudo-labeling, and ground truth (GT) annotations, where GT annotations represent the original PoseSyn method. The Real-only model, trained exclusively on real datasets, is used as the reference model and fine-tuned with each synthesized dataset. MPJPE and PA-MPJPE metrics are reported across multiple datasets to demonstrate the impact of the synthesis methods on TPE performance.

As shown in Tab. 2, our method demonstrates improved performance compared to a model trained solely on a real dataset (*i.e.*, Real-only). Even the approaches that utilize pseudo-label annotations enhance the generalization performance of TPE trained only with real data. This result highlights PoseSyn's potential for leveraging a broader range of unlabeled images, which could significantly expand the diversity of human appearances and challenging poses. Furthermore, the performance improvements observed in 2D pose pseudo-labeling approaches are superior to those in 3D pose pseudo-labeling. This discrepancy arises due to the added complexity of 3D pose pseudo-labeling, which considers depth in three-dimensional space, potentially reducing accuracy when identifying problematic poses in the EEM framework.

Finally, to demonstrate that our method can even leverage various datasets beyond the MPII dataset, we showcase our synthesized dataset which incorporates DeepFashion [20] dataset for reference images and UCF101 [21] dataset for challenging poses, as shown in Fig. 2. The DeepFashion dataset features humans with diverse appearances, making it an excellent choice for reference images. On the other hand, the UCF101 dataset consists of various in-the-wild human poses, making it attractive for use as challenging poses.

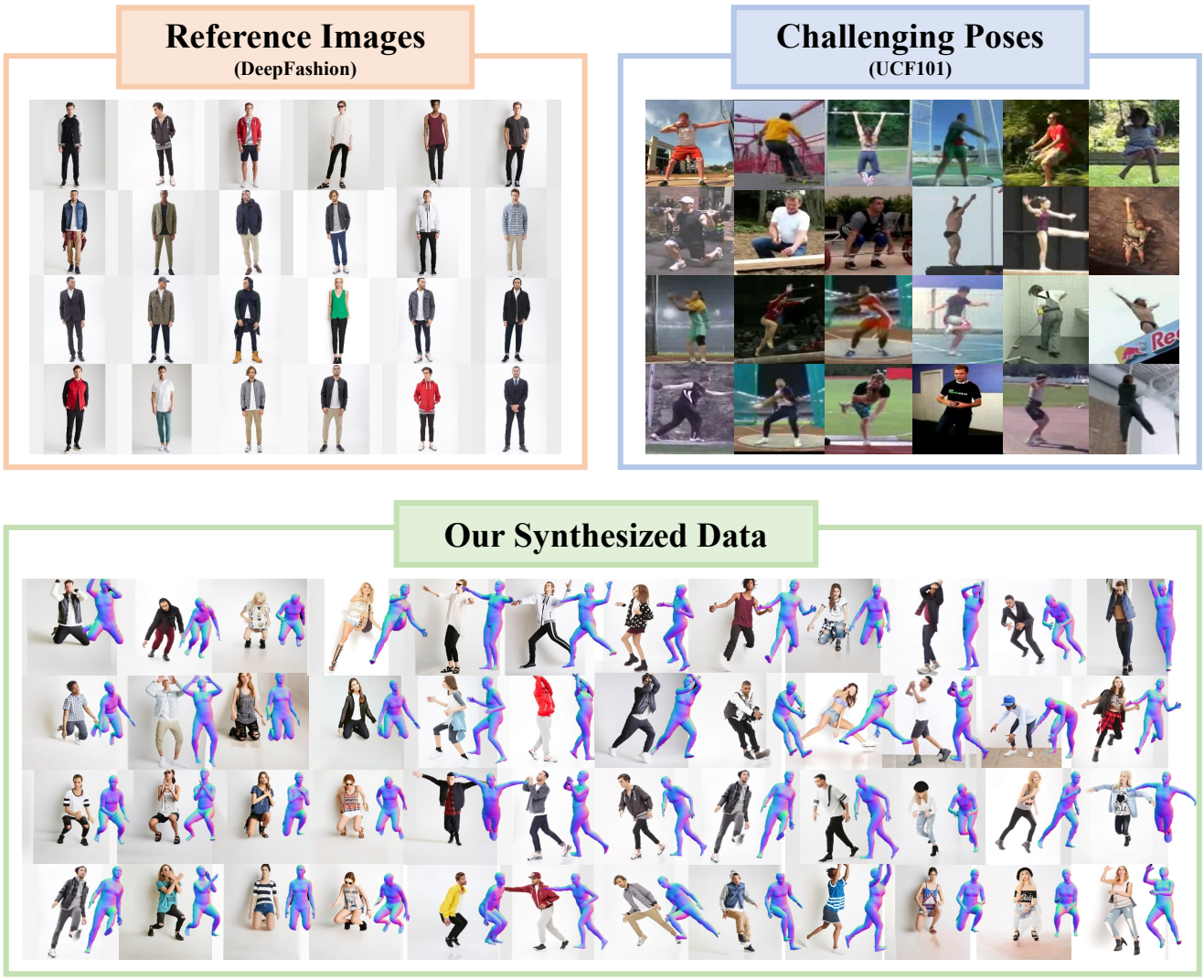
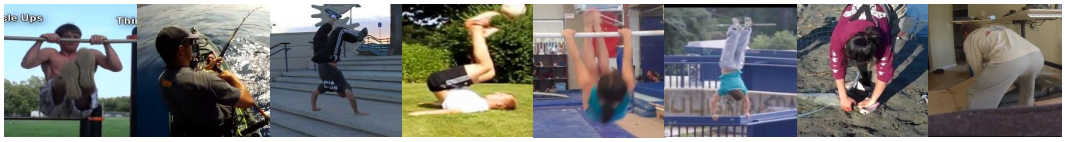


Figure 2. **Our synthesized Data.** We demonstrate the extendibility of our approach in augmenting any easily accessible in-the-wild unlabeled images into a 3D pose dataset. Here, we incorporate DeepFashion dataset (as reference images) and UCF101 dataset (as challenging poses) to create our synthesized 3D pose dataset.

C. More Results

Reference Images in Motion-guided Video Generation We utilize the non-challenging images identified through EEM as reference images in the Motion-guided Video Generation stage. Using challenging images as reference can hinder the motion-guided generative model's [13] ability to perform parametric shape alignment in human animation. This misalignment significantly degrades the quality of the animated images, as shown in Fig. 3.



(a) Challenging images

(b) Generated images

Figure 3. **Experiments on Reference Image in Motion-guided Video Generation.** The first row displays the challenging images identified through EEM, while the second row illustrates the generated images when these challenging images are served as reference images for the human animation model.

Filtering Stage Even when using non-challenging images as references, the motion-guided video generation model occasionally leads to images with artifacts, such as blending humans with the background or missing joints. We present the examples of images removed through our filtering process in Fig. 4.

(a) Background blending

(b) Vanishing limb

Figure 4. **Necessity of Filtering Stage.** The human animation model occasionally generates images with visual artifacts, including either (a) the unnatural blending of human figures with their backgrounds or (b) disappearance of some joints. These issues highlight the necessity of the filtering stage in our framework. All presented images are examples of filtered-out samples during the stage.

Diversity in Synthesis Our method offers promising advantage of augmenting a single reference image through various challenging poses, as well as augmenting a single problematic pose through various reference images with different viewpoints, human appearances, and backgrounds. This dual capability allows for the creation of a highly diverse dataset. We visualize the synthesized data samples in Fig. 5 and Fig. 6.

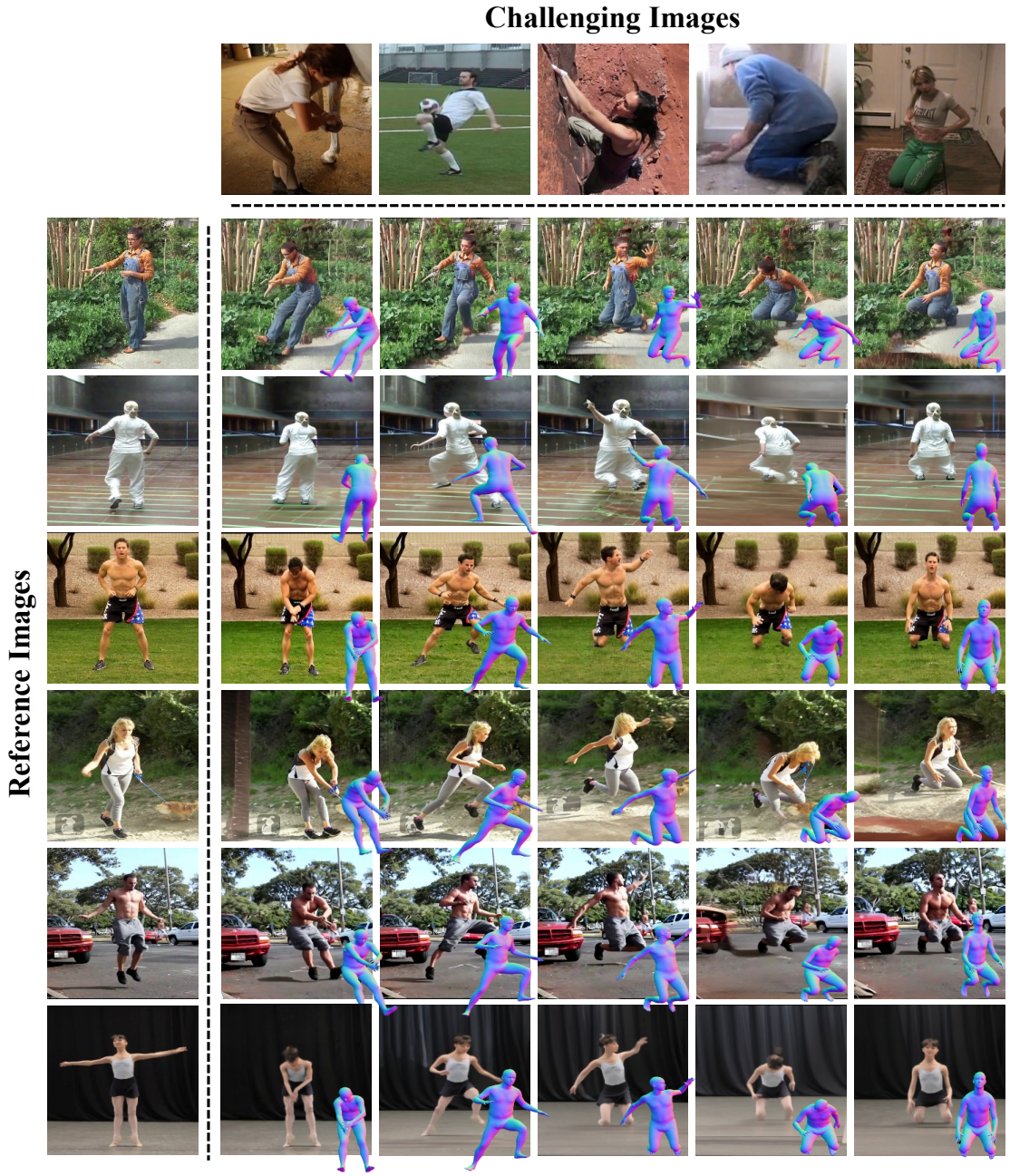


Figure 5. **Visualization of Our Synthesized Dataset with Various Combinations.** The first column on the far left displays non-challenging images identified through EEM, which are served as reference images. On the other hand, the top row presents challenging images with problematic poses, also identified through EEM, where these poses are augmented into challenging motion sequences via MSM. The remaining images are synthesized data samples created by combining the non-challenging images with the challenging poses, resulting in diverse datasets with varied appearances, poses, and backgrounds.

Figure 6. **Visualization of Our Synthesized Dataset.** Our framework synthesizes dataset consisting of highly diverse image sets, featuring various backgrounds, human appearances, and challenging poses.

D. Additional Qualitative Results

We present additional qualitative comparison results using three types of TPEs (*i.e.*, 3DCrowdNet, Hybrik, and Vanilla) in Fig. 7, Fig. 8, and Fig. 9.

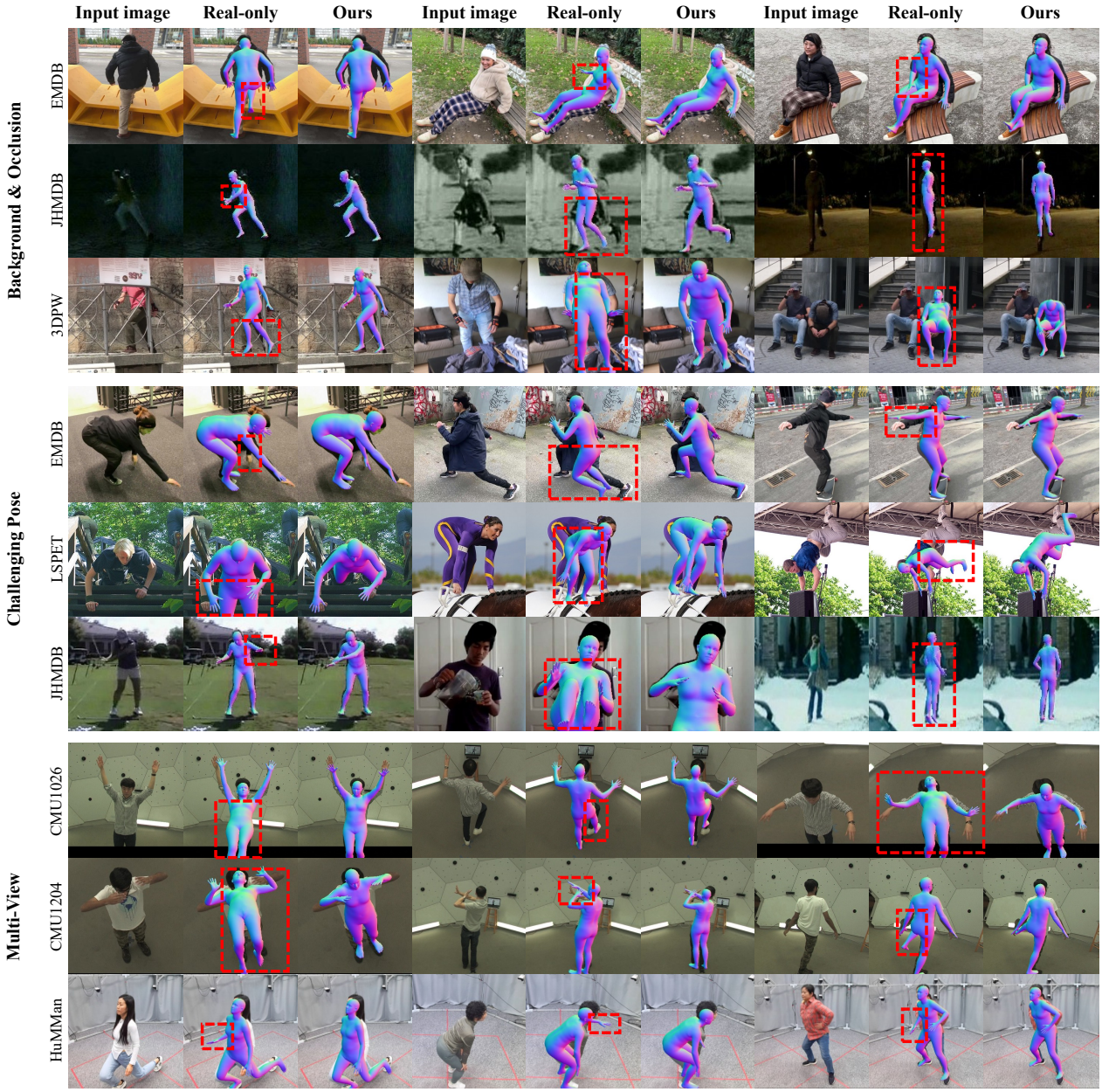


Figure 7. **Qualitative Results (3DCrowdNet).** Our approach improves generalization of 3DCrowdNet trained solely on real dataset (*i.e.*, Real-only) across various benchmarks. Red boxes highlight areas of incorrect predictions in the model trained with real-only data.

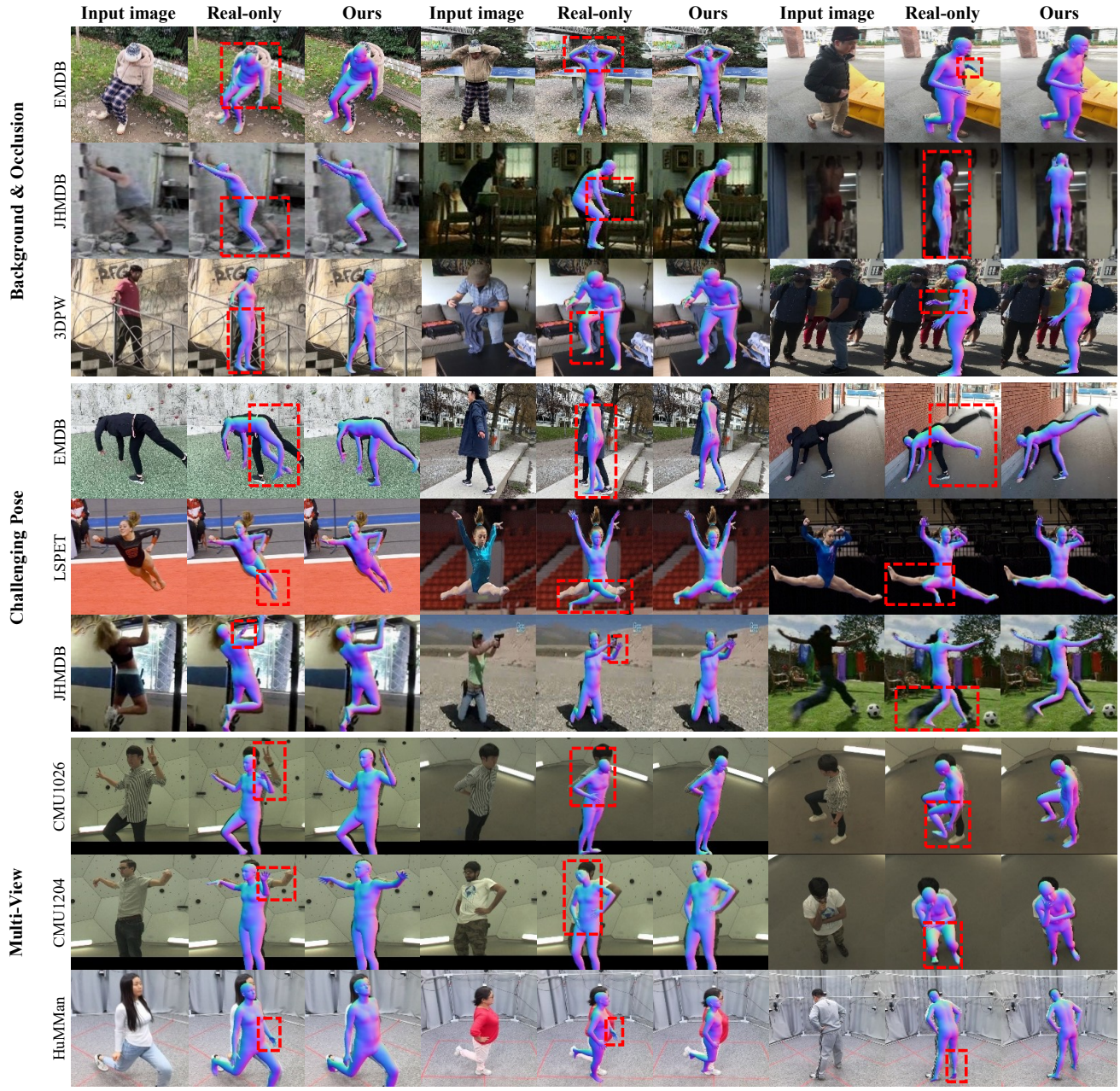


Figure 8. **Qualitative Results (Hybrik).** Our approach improves generalization of Hybrik trained solely on real dataset (*i.e.*, Real-only) across various benchmarks. Red boxes highlight areas of incorrect predictions in the model trained with real-only data.

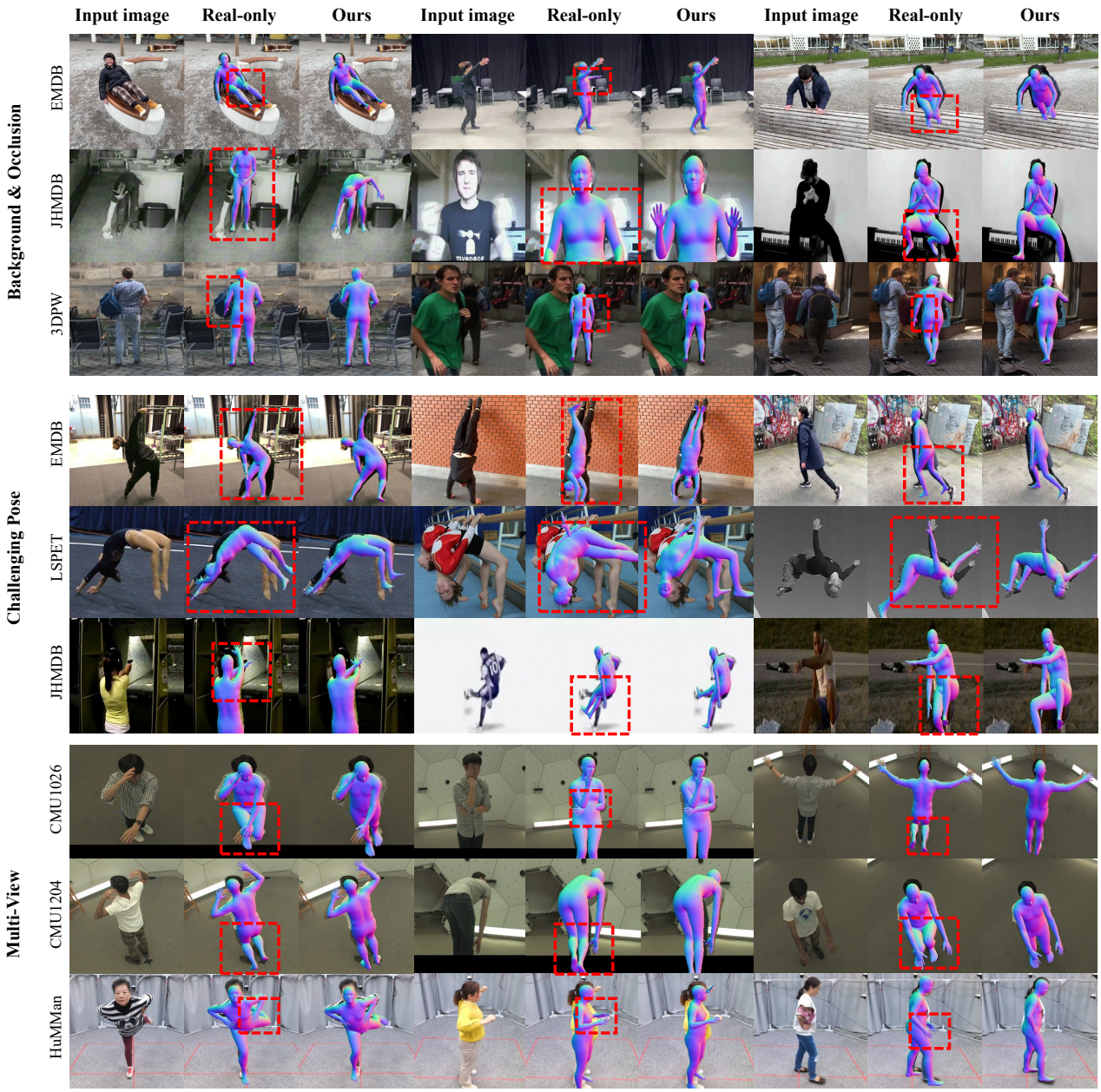


Figure 9. **Qualitative Results (4DHumans).** Our approach improves generalization of Vanilla trained solely on real dataset (*i.e.*, Real-only) across various benchmarks. Red boxes highlight areas of incorrect predictions in the model trained with real-only data.

References

[1] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 1, 3

[2] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1475–1484, 2022. 1, 3

[3] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 1, 3

[4] 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, jul 2014. 1

[5] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 1

[6] 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. 1

[7] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 4

[8] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 1

[9] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 1

[10] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 1

[11] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 1

[12] Weizhi Wang. Llava-llama-3-8b: A reproduction towards llava-3 based on llama-3-8b llm backbone, 2024. 1

[13] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 6

[14] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[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 (CVPR)*, pages 5152–5161, June 2022. 3

[17] 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. 3

[18] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 4

[19] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. Multi-person 3d pose and shape estimation via inverse kinematics and refinement. In *ECCV*, 2022. 4

[20] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5

[21] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 5