

Supplementary Materials: Generative Modeling of Shape-Dependent Self-Contact Human Poses

1. Dataset Details

Data capture protocol: Our dataset is constructed on the minimally-clothed body setup of [6], which aligns with the previous work on body shape prior captured with subjects in tight-fit clothing [5, 9]. We obtain user consent for data captures and the release of the registered SMPL-X parameters.

Due to the fixed capture space, most samples do not have high variations for lower-body poses. Nevertheless, this setup allows for capturing *intricate* upper-body self-contact details (e.g., “rubbing eyes”) with unprecedented fidelity unavailable in existing studies [1, 2, 7, 10]. While modeling large variations in lower-body pose (e.g., tying shoelaces) is not prioritized in this work, we will consider an expanded capture setup as future work.

SMPL-X registration: We initiate with a human mesh model used in [6] that has a uniform topology across subjects, and we pre-compute its vertex-face correspondence to SMPL-X using barycentric coordinates. We register the human model across frames while tracking pose and surface precisely without relying on mocap markers. Given multi-view dome captures, we first fit the human model to the rest pose (A-Pose). Then we run 3D pose tracking based on multi-view images over the frames and use Linear Blend Skinning (LBS) that transforms a mesh in the rest pose to the desired pose of each frame. The subject’s poses are continuously captured at 30 Hz with scripted action instructions to let participants express the corresponding gestures. Given the registered mesh, the SMPL-X registration is obtained through vertex-to-vertex alignment between two meshes¹, as shown in Fig. 6. The continuous poses in our capture allow stable mesh alignment by using the previous frame’s registration as initialization for the current frame, preventing significant fitting failures.

Data illustration: Additional supplementary videos are included. Videos 1-1 and 1-2 show our captured data, registered meshes, and contact maps from the Goliath-4’s subjects [6]. Video 2 illustrates more captured sequences. Video 3 is a video of the contact heatmap. These include fine self-contact interactions with high-fidelity mesh registration. The heatmap suggests a high contact likelihood on hands and across body parts, e.g., face, neck, belly, arm,

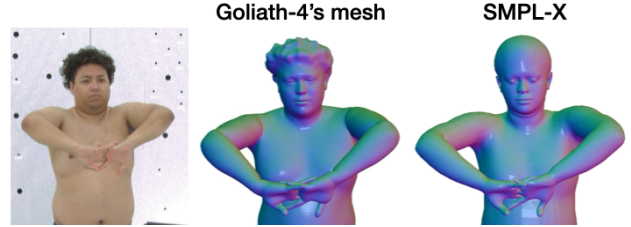


Figure 6. Conversion from Goliath-4 [6]’s mesh to SMPL-X.

back, and thigh.

Data statistics: Fig. 7 details action scripts used in the capture and the number of self-contact poses per action. The nouns of the actions suggest interacting body parts as following groups.

- Head-related: Face, forehead, temples, eyes, nose, hair, facial hair, and neck
- Upper body-related: Arm, hand, wrist, fingers, thumb, and palm
- Torso-related: Belly, back, lumbar, and thighs

Instead, the verbs indicate how to interact with the body part; general movements are represented, such as hitting, grabbing, holding, clapping, rubbing, massaging, scratching, punching, wrapping, and itching. In addition, hand-specific movements include extension, flex, rotation, press, snap, interlock, touch, and squeeze. These hand interactions tend to be in close contact mostly in the captured sequence, resulting in a large number of poses in self-contact, such as “hands massaging hands”.

Our dataset is constructed by capturing 130 subjects where the gender distribution is detailed in Tab. 1. To confirm the variety of captured shape information, we provide comprehensive analysis on shape statistics in Fig. 8, i.e., standard deviation and range of 10 shape components of SMPL-X compared to the existing self-contact datasets, such as HumanSC3D [2] and MTP [7]. Our dataset (red) has the largest variety in most components except for 6th range, 8th std and range, indicating higher subject diversity and variability of our Goliath-SC.

2. Additional Implementation Details

Baselines: We detail the implementation of baselines used in our experiments. For fair comparison, we retrain the

¹https://github.com/vchoutas/smplx/blob/main/transfer_model/docs/transfer.md

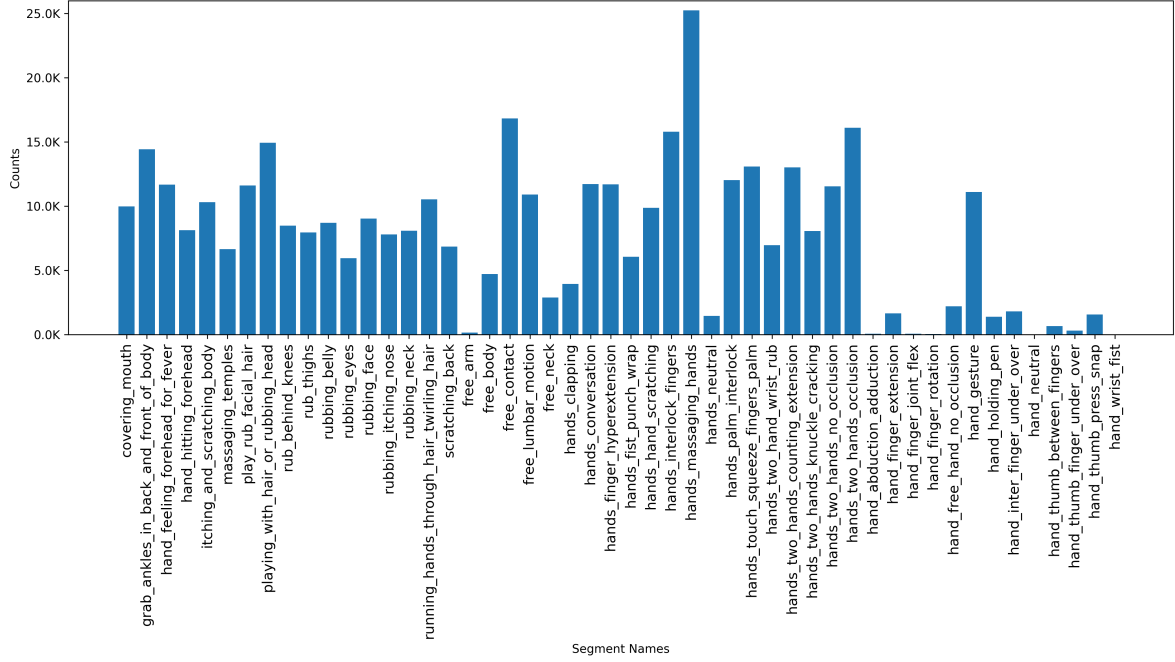


Figure 7. **Statistics of scripted actions and the number of self-contact poses in Goliath-SC.**

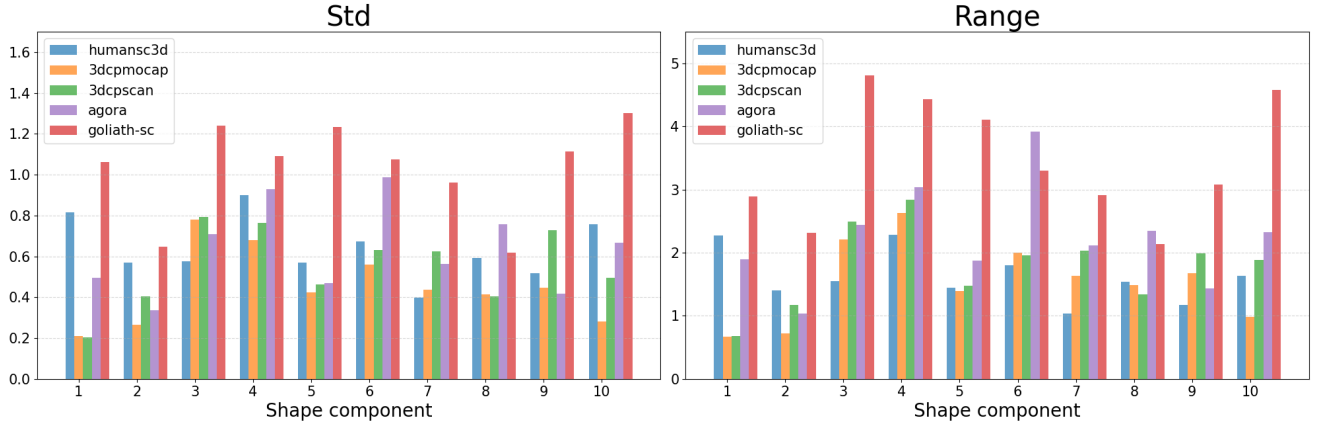


Figure 8. **Variability of subject shapes.** Standard deviation and range (max — min) of the first 10 shape components in self-contact datasets, namely HumanSC3D [2], (3DCPMocap, 3DCPScan, Agora) from MTP [7], and our Goliath-SC.

comparison models from scratch with the same input representation of the whole-body pose parameters (aligned to \mathbf{X} of Sec. 4.2), including hands, body, and face.

BUDDI [8] is originally proposed for two-body interactions, modeling the joint distribution of the body pose parameters of SMPL-X (compatible to SMPL) and its shape parameters without latent diffusion modeling. A two-hand interaction generation model, InterHandGen [4], shares a similar architecture. Our implemented BUDDI* modifies the original BUDDI to take the whole-body pose parameters of a single person. Following the original implementation,

the transformer layers are used and all pose parameters are concatenated to a single vector, which indicates the absence of part-wise attention compared to our PAPoseDiff. To produce the results of Tab. 2, we construct baselines of the joint distribution modeling between pose and shape, *i.e.*, unconditional model, and shape-conditional pose generation with the input embedding of the shape parameters. In addition, when adapting to the task of single-view refinement, we use our fitting algorithm (Algorithm 1) with the unconditional BUDDI* prior. The hyper-parameters in the fitting (*e.g.*, weight for 2D keypoint fitting and start diffusion time) fol-

low those of our final method.

VPoser [9] is a VAE-based pose prior that learns pose distribution on the body pose parameters of SMPL-X. Similarly to BUDDI*, we adapt this architecture for our task, by taking the whole-body pose parameters and adding shape parameters as conditional input, denoted as VPoser*.

Training details: We train generative models for 150,000 iterations with a batch size of 32, using an Adam optimizer [3] with a learning rate of $1e-4$. Our diffusion process is based on cosine noise scheduling with $T=1000$. The auto-encoder network consists of a single linear layer with a hidden size of 256, which has separate weights for each body part. Unlike the conventional choice of using 10 shape components of SMPL-X, we input the full 300-dimensional vector of the shape parameters to let the generative prior access as much fine details as possible (e.g., hands and face shapes).

3. Additional Results

Qualitative results of shape interpolation: Video 4 shows additional results of shape interpolation with our shape-conditional PAPoseDiff. Poses are sampled from interpolated shape parameters and fixed noise input, and four self-contact videos are concatenated into the Video 4. The results indicate that different noise inputs correspond to non-identical self-contact patterns. With the shape changes, we find that the interacting parts are almost consistent and no significant corruption is observed. This suggests that the proposed diffusion prior enables learning a smooth pose manifold dependent on the given shapes.

Qualitative results of single-view pose estimation: Additional qualitative results of single-view pose estimation are found in Fig. 9, including SMPLer-X, fine-tuned SMPLer-X[†], 2D fitting, BUDDI*, our final refinement (Ours), and GTs. We observe that hands are not often in contact with SMPLer-X (e.g., Rows 1,2,4,6), while the fine-tuned baseline struggles with highly bent hand fingers (e.g., Rows 3,6,7) and incorrect contact states, e.g., for the hidden left hand behind the neck of Row 5. The 2D keypoint fitting baseline tends to exhibit unsolved depth ambiguity (Rows 1,6) and implausible hand poses (Row 1) due to over-fitting to the 2D observation. The BUDDI* method often relies heavily on the model prior with a large 2D error to the observation. This indicates that the method generates plausible poses, yet not aligned to the given 2D keypoints, such as Rows 1,2,3,5. It also comprises higher hand depth errors (to those to be in contact) like Rows 1,2,6. Notably, our method can resolve these failures presented by the comparison models and shows significantly reduced errors in 3D compared to the GT.

The last row shows a remaining failure when fingers are in complex interaction, i.e., the fingers of both hands, except for the ring fingers, are overlapped while only the ring

fingers are bent. Neither method handles this pose well because of the inaccuracy of 2D keypoint detection. Improving detection and model-based refinement to such fine interactions are future challenges.

References

- [1] M. Fieraru, M. Zanfir, E. Oneata, A. Popa, V. Olaru, and C. Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7212–7221, 2020. 1
- [2] M. Fieraru, M. Zanfir, E. Oneata, A. Popa, V. Olaru, and C. Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1343–1351, 2021. 1, 2
- [3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 3
- [4] J. Lee, S. Saito, G. Nam, M. Sung, and T. Kim. Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 527–537. IEEE, 2024. 2
- [5] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics (ToG)*, 34(6):248:1–248:16, 2015. 1
- [6] J. Martinez, E. Kim, J. Romero, T. Bagautdinov, S. Saito, S.-I. Yu, S. Anderson, M. Zollhöfer, T.-L. Wang, S. Bai, S.-E. Wei, R. Joshi, W. Borsos, T. Simon, J. Saragih, P. Theodosis, A. Greene, A. Josyula, S. M. Maeta, A. I. Jewett, S. Venshtain, C. Heilman, Y.-T. Chen, S. Fu, M. E. A. Elshaer, T. Du, L. Wu, S.-C. Chen, K. Kang, M. Wu, Y. Emad, S. Longay, A. Brewer, H. Shah, J. Booth, T. Koska, K. Haidle, J. C.-H. Hsu, T. Dauer, P. Selednik, T. Godisart, S. Ardisson, M. Cipperly, B. Humberston, L. Farr, B. Hansen, P. Guo, D. Braun, S. Krenn, H. Wen, L. Evans, N. Fadeeva, M. Stewart, G. Schwartz, D. Gupta, G. Moon, K. Guo, Y. Dong, Y. Xu, T. Shiratori, F. A. Prada Nino, B. R. Pires, B. Peng, J. Buffalini, A. Trimble, K. A. A. McPhail, M. R. Schoeller, and Y. Sheikh. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*. IEEE, 2024. 1, 5
- [7] L. Müller, A. A. A. Osman, S. Tang, C. P. Huang, and M. J. Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999. Computer Vision Foundation / IEEE, 2021. 1, 2
- [8] L. Müller, V. Ye, G. Pavlakos, M. J. Black, and A. Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9687–9697. IEEE, 2024. 2
- [9] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1, 3

- [10] Y. Yin, C. Guo, M. Kaufmann, J. J. Zarate, J. Song, and O. Hilliges. Hi4D: 4D instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17016–17027, 2023. [1](#)



Figure 9. **Qualitative results of single-view pose estimation.** The four subjects of Goliath-4 [6] are illustrated.