Supplementary Material
One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer

Jing Lin[1,2§], Ailing Zeng[1¶], Haoqian Wang[2], Lei Zhang[1], Yu Li[1]
[1] International Digital Economy Academy (IDEA),
[2] Shenzhen International Graduate School, Tsinghua University
https://osx-ubody.github.io

Overview

This supplementary material presents more details and additional results not included in the main paper due to page limitation. The list of items included are:

- More experiment setup and details in Sec. A.
- Efficiency comparison with SOTA in Sec. B.
- Experiment on AGORA dataset in Sec. C.
- More introduction of UBody in Sec. D.
- Inter-scene benchmark on UBody dataset in Sec. E.
- Qualitative comparisons with SOTA in Sec. F.

A. Experiment Setup

Evaluation metrics. To quantitatively evaluate the performance of human mesh recovery, MPVPE, PA-MPVPE, MPJPE, and PA-MPJPE are used as evaluation metrics. Besides, we also report normalized mean vertex error (NMVE) and normalized mean joint error (NMJE) by the standard detection metric, F1 score (the harmonic mean of recall and precision) to penalize models for misses and false positives on AGORA test set with many multi-person scenes.

Implementation details. Our *OSX* model is implemented in Pytorch. It is trained with Adam optimizer ($\beta_1 = 0.1, \beta_2 = 0.999$) using the Cosine Annealing scheme for 14 epochs. The learning rate is initially set to 1×10^{-4}. The batch size is set to 192. Random scaling, rotation, horizontal flip, and color jittering are used as data augmentations during training. The spatial size of the input image is 256×192. The number of body tokens $\mathbf{T}_b$ and component tokens $\mathbf{T}_c$ are set to 27 and 92, respectively. During experiments on the AGORA-test set, we remove the decoder as we find that the decoder increases training time and does not significantly improve performance on AGORA-test set. This observation may be attributed to the fact that the main problem of AGORA is occlusion, while the decoder aims to estimate hands/face at a finer level.

B. Efficiency comparison with SOTA methods

We report the complexity comparisons including average inference time, number of model parameters, FLOPs, and the NMJE-All on AGORA-test in Table S-1. The numbers are measured for single-person regression on the same resolution input using a machine with an NVIDIA A100 GPU. OSX has *the shortest inference time and lowest error*, indicating the advantages in practical applications.

Method	ExPose [34]	PIXIE [13]	H4W [27]	PyMAF-X [48]	OSX
NMJE-All (mm)	263.3	230.9	141.1	140.0	127.6
Infer Time (ms)	120.2	192.0	73.3	209.3	54.6
Params (M)	135.8	192.9	77.9	205.9	102.9
FLOPS (G)	28.5	34.3	16.7	35.5	25.3

Table S-1. Efficiency comparisons with multi-stage methods.

C. Experiment on AGORA Dataset

In this part, we report the complete result on the AGORA test set and the experiment result on the AGORA val set.

AGORA Test Set. Table S-2 depicts the complete result on the AGORA test set. All the results are taken from the official leaderboard. As shown, our OSX outperforms other competitors on most metrics, especially on the evaluation of the body and full-body recovery. More specifically, for full-body reconstruction, OSX even surpasses PyMAF-X [33] by 10.6 mm, 9.1 mm, 2.9 mm, and 4.7 mm on NMVE, NMJE, MVE, and MPJPE, respectively. Since PyMAF-X has a lower detected person ratio, they have similar results on MVE and MPJPE metrics, which only calculate the matched person. The NMVE and NMJE will take the misses and false positives into account, and we have overall better multi-person estimation with more improvement under the metrics. Notably, although OSX does not use extra

§ Work done during an internship at IDEA; ¶ Corresponding author.

Method	NMVE ↓		NMJE ↓		MVE ↓				MPJPE ↓			
	Full-Body	Body	Full-Body	Body	Full-Body	Body	Face	LH/RH	Full-Body	Body	Face	LH/RH
SMPLify-X [20]	333.1	263.3	326.5	256.5	236.5	187.0	48.9	48.3/51.4	231.8	182.1	52.9	46.5/49.6
ExPose [19]	265.0	184.8	263.3	183.4	217.3	151.5	51.1	74.9/71.3	215.9	150.4	55.2	72.5/68.8
FrankMocap [26]	-	207.8	-	204.0	-	168.3	-	54.7/55.7	-	165.2	-	52.3/53.1
PIXIE [4]	233.9	173.4	230.9	171.1	191.8	142.2	50.2	49.5/49.0	189.3	140.3	54.5	46.4/46.0
Hand4Whole [16] †	144.1	96.0	141.1	92.7	135.5	90.2	41.6	46.3/48.1	132.6	87.1	46.1	44.3/46.2
PyMAF-X [33] †	141.2	94.4	140.0	93.5	125.7	84.0	**35.0**	**44.6/45.6**	124.6	83.2	**37.9**	**42.5/43.7**
OSX (Ours) †	**130.6**$_{\downarrow 7.5\%}$	**85.3**$_{\downarrow 9.6\%}$	**127.6**$_{\downarrow 8.9\%}$	**83.3**$_{\downarrow 10.9\%}$	**122.8**	**80.2**	36.2	45.4/46.1	**119.9**	**78.3**	**37.9**	43.0/43.9

Table S-2. Reconstruction errors on the AGORA test set. † denotes the methods that are fine-tuned on the AGORA training set or similarly synthetic data [11]. The best results are shown in **bold** and the second best results are highlighted with underlined font.

hand-only and face-only datasets, it can achieve competitive results on hand and face metrics, which demonstrates the effectiveness of our component-aware decoder.

AGORA Val Set. Table S-3 shows the result on the AGORA val set. All the results are taken from [16] except OSX. Although we do not use extra hand/face specific datasets during training, OSX outperforms the SOAT method Hand4Whole by 8.3% on the MPVPE-all, demonstrating the effectiveness of our one-stage method.

Method	MPVPE ↓			PA-MPVPE ↓		
	All	Hand	Face	All	Hand	Face
ExPose [19]	219.8	115.4	103.5	88.0	12.1	4.8
FrankMocap [26]	218.0	95.2	105.4	90.6	11.2	4.9
PIXIE [4]	203.0	89.9	95.4	82.7	12.8	5.4
Hand4Whole [16]	183.9	72.8	81.6	73.2	**9.7**	**4.7**
OSX (Ours)	**168.6**$_{\downarrow 8.3\%}$	**70.6**	**77.2**	**69.4**	11.5	4.8

Table S-3. Reconstruction errors on the AGORA val set.

D. UBody: An Upper Body Dataset

D.1. Data Collection

To bridge the gap between the basic human mesh recovery task and its downstream applications, we design *UBody* with two rules. First, we research a wide range of human-related downstream tasks with upper-body scenes, including gesture recognition [7, 15, 30], sign language recognition, and translation [2, 3, 10, 24, 28, 34], person clustering [1], emotion analysis, speaker verification [17], micro-gesture understanding [14], audio-visual generation and separation [23], human action recognition, and localization [5, 6, 21, 25, 27], and human video segmentation [12]. We select the corresponding high-quality datasets from these existing tasks as a part of our data for the corresponding scenarios. In order to ensure a balanced amount of data for each scene, for datasets with many videos (*e.g.*, lasting 20k minutes), we manually selected the videos in which the upper body appeared more frequently. Second, with all kinds of athletic competitions, entertainment shows, we media, online conferences, and online classes being more and more indispensable, we carefully selected a large number of rich videos from YouTube to provide new opportunities and challenges for potential applications. Since some untrimmed videos may have missing main characters, extraneous images such as opening and closing credits, and repetitive actions, we manually fine-cut the long videos. Each edited video is 10 seconds long, which ensures the high quality of the video. In order to prevent infringement of ownership rights, we only provide download links to the corresponding videos and our labels without any personal information. In summary, we collect fifteen real-life scenarios with more than **105,1k** frames. We split the train/test sets from two protocols as follows.

- *Intra-scene*: in each scene, the former 70% of the videos are the training set, and the last 30% are the test set. The benchmark was provided in the main paper.
- *Inter-scene*: we use ten scenes of the videos as the training set and the other five scenes as the test set. Due to the page limit, we present the benchmark in Table S-4.

D.2. Data Annotation Processes

As shown in Figure S-1, we design a thorough whole-body annotation pipeline with high precision. It is divided into two stages: 2D whole-body keypoint annotation and 3D SMPLX annotations fitting. Since *UBody* scenes have a number of unpredictable transitions and cutscenes that make it difficult to use the temporal smoothing approaches [22,31,32], the annotation is conducted on a single frame.

2D whole-body keypoint annotation: We first detect all persons and their hands in an image via a specific human and hand detector *BodyHands* [18] shown as *Body Detector* and *Hand Detector* in Figure S-1. Leveraging the recent state-of-the-art 2D pose estimator *ViT-Body-only* [29], we use the pre-trained model trained on the COCO [13] dataset to localize 17 body keypoints for each detected single person, named K_{Body}, which shows highly robust results on many scenes. Due to the diverse scales and motion blur for the fast-moving hands, we find that *Hand Detector* will output false positive samples or miss some hands. To enhance the performance of hand detection, we train a 2D

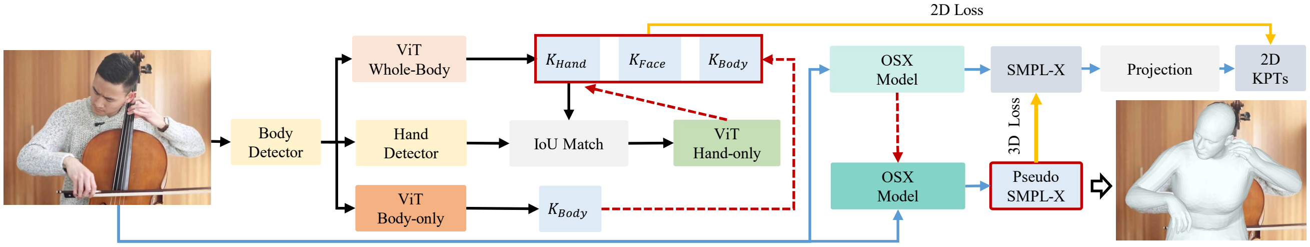


Figure S-1. Illustration of the annotation pipeline of *UBody*. Black lines show the annotation process of 2D whole-body keypoints, and blue lines are the 3D SMPL-X annotation procedure. Red dotted lines mean to update the information.

whole-body estimator on COCO-wholeBody [9] with 133 2D keypoints, called *ViT-WholeBody* following the model design of ViTPose [29] and masked autoencoder pre-trained scheme [8]. *ViT-WholeBody* can provide high-recall hand keypoints K_{Hand}, but the localization precision is low because of the fully one-stage pipeline and low-resolution of hands from the raw image. Accordingly, We can obtain coarse hand bounding boxes by calculating the maximum, and minimum values of the detected left and right-hand keypoints to correct the hand boxes from *Hand Detector* via an IoU matching strategy. Then, we use the fine hand boxes to crop the hand patches, resize them to a larger size, and put them into our specific pre-trained *ViT-Hand-only* model trained with the hand labels from the COCO-Whole dataset. In summary, *ViT-WholeBody* will output the body, hand, and face 2D keypoints. We use the body output from *ViT-Body-only* to replace the K_{Body}, and use the fine hand keypoints from *ViT-Hand-only* to change the K_{Hand}. As the face of the current SMPL-X model does not require much detail, we simply use the 2D face keypoints K_{Face} obtained from *ViT-WholeBody*.

3D whole-body mesh recovery annotation: Different from previous optimization-based annotation [20] that may output implausible poses, we use our proposed *OSX* to estimate the SMPL-X parameters from human images as a proper 3D initialization to provide pseudo-3D constraints. Benefiting from current 2D keypoint localization that tends to be more accurate, we additionally supervise the projected 2D whole-body keypoints by the above annotated 2D whole-body keypoints as a way to train *OSX*. More importantly, to avoid performance degradation from not accurate enough initial labeling and consistently push up the 3D annotation quality, we propose an iterative training-labeling-revision loop for every 30 epochs to train 120 epochs in total.

E. Inter-Scene Benchmark on UBody dataset

Due to the page limit, we further provide another data protocol comparison to show the usage of the proposed *UBody*. Table S-4 presents the performance comparisons of existing 3D whole-body methods. Inter-scene test shows large errors than the intra-scene test due to the different motion and gesture distributions. The model finetuned on AGORA still has a significant gap than trained on the COCO dataset. Furthermore, we also train Hand4Whole and *UBody* on our training set, we can find a consistent improvement compared to the original pretrained model, indicating that *UBody* can serve to bridge the gap among these downstream real-life scenes. Moreover, different from single-frame AGORA and EHF, *UBody* provides videos, which can drive progress in spatial-temporal modeling on such edit media sources.

Method	MPVPE ↓			PA-MPVPE ↓		
	All	Hand	Face	All	Hand	Face
ExPose [19]	185.7	89.5	47.2	76.4	11.8	4.0
PIXIE [4]	185.0	60.9	45.3	74.5	11.9	4.2
Hand4Whole [16]×	198.1	66.9	51.8	90.2	10.3	4.1
Hand4Whole [16]	109.4	50.4	24.8	57.0	8.9	2.7
Hand4Whole [16]†	87.4	**41.6**	22.1	46.3	**8.0**	2.0
OSX (Ours)	100.7	52.5	24.5	52.9	9.5	2.6
OSX (Ours)†	**82.0**	44.2	**21.5**	**44.2**	8.8	**1.9**

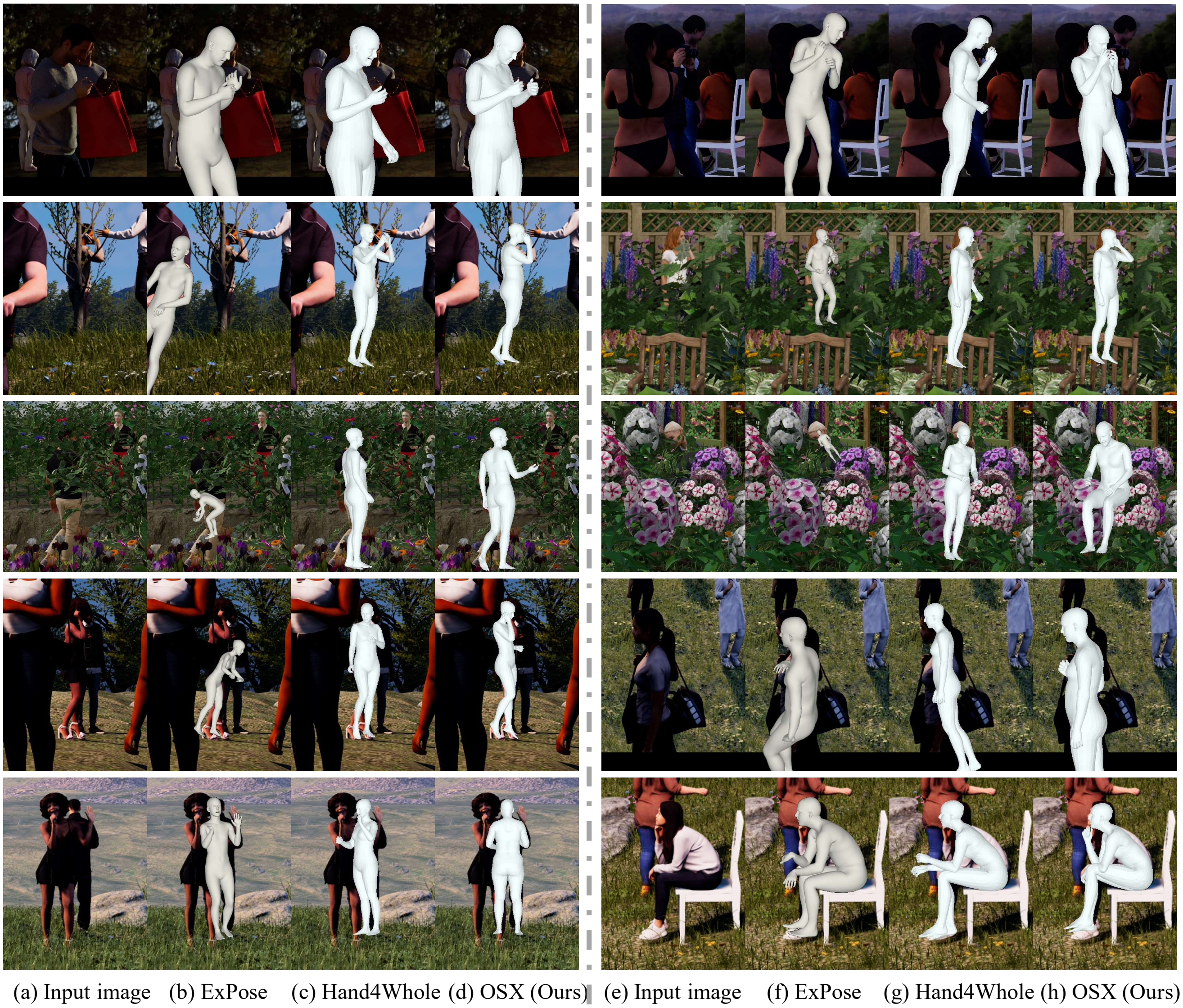
Table S-4. Reconstruction errors on *UBody* test set on the *inter-scene* protocol. All models are pretrained on previous datasets, except for the results labeled by (i) †: finetuned on the *UBody* training data; (ii) ×: finetuned on the AGORA training data.

F. Qualitative with SOTA method

Qualitative comparisons on AGORA: We compare the mesh quality on the AGORA dataset in Figure S-2. Agora is a synthetic dataset with many challenging factors like heavy occlusion, dark environment, and unnatural multi-person interaction. It only has limited actions, *e.g.*, taking phones, walking, sitting, *etc*. We can see *OSX* outperforms ExPose [19] and Hand4Whole [16] consistently in terms of global body orientations, whole-body poses, and hand pose.

Qualitative comparisons on EHF: The visual comparisons of whole-body mesh recovery quality on the EHF dataset can be found in Figure S-3. As can be seen, *OSX* estimates the most accurate whole-body poses, in which the body parts like hands, feet, and hands are better aligned with the person in the image.

Qualitative comparisons on UBody: The qualitative com-



(a) Input image (b) ExPose (c) Hand4Whole (d) OSX (Ours) (e) Input image (f) ExPose (g) Hand4Whole (h) OSX (Ours)

Figure S-2. Comparisons of existing 3D whole-body estimation methods on AGORA.

parison on our *UBody* is in Figure S-4. *UBody* focuses more on the expressive upper body part. Hand4Whole [16] and our *OSX* produces better body mesh recoveries than ExPose [29]. Close inspection of the hand part shows that our hand recovery is more accurate than Hand4Whole.

Visualization of our annotation on UBody: The visualizations of our SMPL-X annotation in our *UBody* can be found in Figure S-5, S-6, and S-7. Our annotation produces high-quality ground truth. In many challenging cases of expressive hand poses, our estimated mesh can capture fine-level details.

References

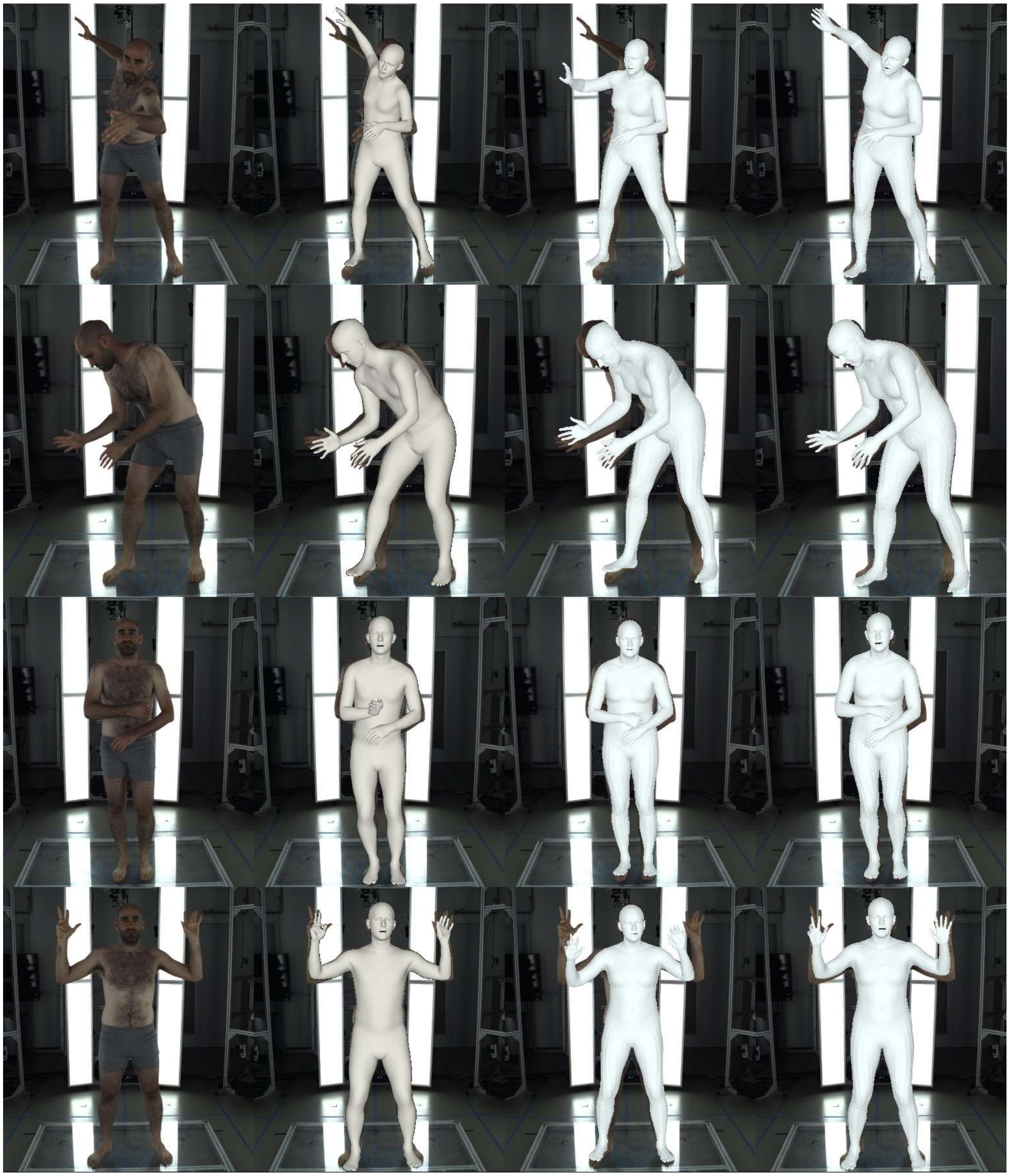
[1] Andrew Brown, Vicky Kalogeiton, and Andrew Zisserman. Face, body, voice: Video person-clustering with multiple modalities. In *ICCV*, 2021. 2

[2] Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. Content4all open research sign language translation datasets. In *FG*. IEEE, 2021. 2

[3] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *CVPR*, 2021. 2

[4] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *3DV*, 2021. 2, 3

[5] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018. 2



(a) Input image (b) ExPose (c) Hand4Whole (d) OSX (Ours)

Figure S-3. Comparisons of existing 3D whole-body estimation methods on EHF.

[6] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2

[7] Lin Guo, Zongxing Lu, and Ligang Yao. Human-machine interaction sensing technology based on hand gesture recognition: A review. *IEEE Transactions on Human-Machine Systems*, 2021. 2

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3

[9] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 3

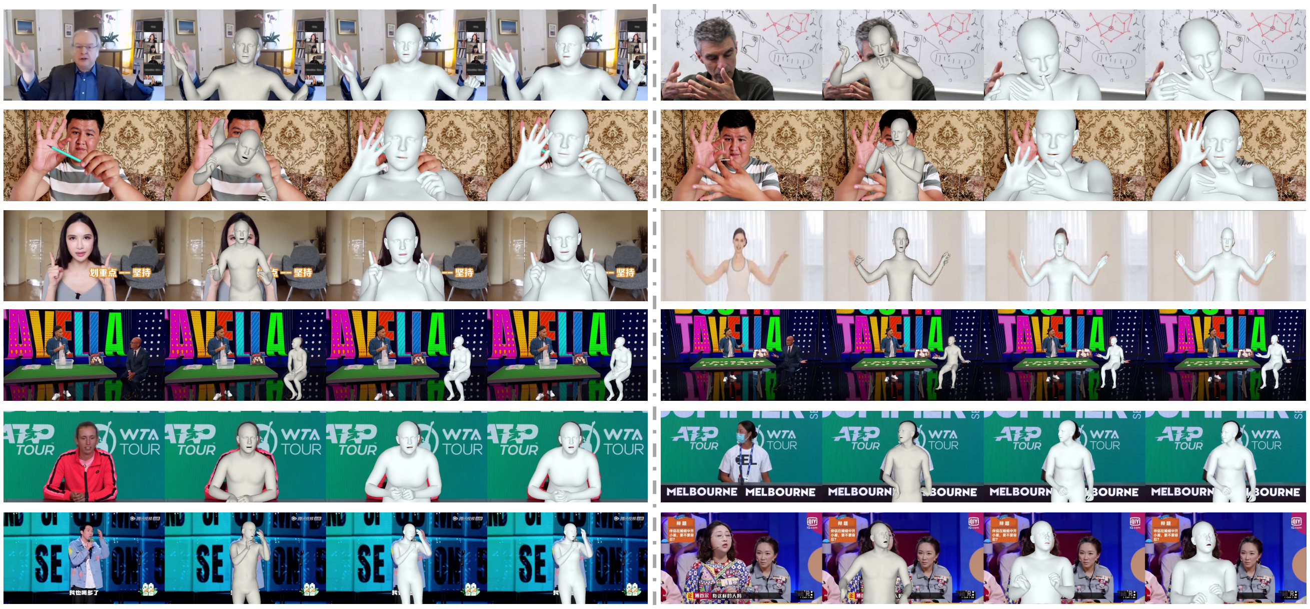


Figure S-4. Comparisons of existing 3D whole-body estimation methods on our proposed *UBody*.

[10] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *ArXiv*, abs/1812.01053, 2019. 2

[11] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *CVPR*, 2021. 2

[12] Zijian Kuang and Xinran Tie. Flow-based video segmentation for human head and shoulders. *arXiv preprint arXiv:2104.09752*, 2021. 2

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[14] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *CVPR*, 2021. 2

[15] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2007. 2

[16] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPRW*, 2022. 2, 3, 4

[17] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 2020. 2

[18] Supreeth Narasimhaswamy, Thanh Nguyen, Mingzhen Huang, and Minh Hoai. Whose hands are these? hand detection and hand-body association in the wild. In *CVPR*, 2022. 2

[19] Georgios Pavlakos, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, Michael J. Black, Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. *ECCV*, 2020. 2, 3

[20] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 2, 3

[21] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *CVPR*, 2022. 2

[22] William H Press and Saul A Teukolsky. Savitzky-golay smoothing filters. *Computers in Physics*, 1990. 2

[23] Jie Pu, Yannis Panagakis, Stavros Petridis, and Maja Pantic. Audio-visual object localization and separation using low-rank and sparsity. In *ICASSP*. IEEE, 2017. 2

[24] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. Sign language production: a review. In *CVPR*, 2021. 2

[25] Chris Rockwell and David F Fouhey. Full-body awareness from partial observations. In *ECCV*, 2020. 2

[26] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCVW*, 2021. 2

[27] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 2

[28] S Subburaj and S Murugavalli. Survey on sign language recognition in context of vision-based and deep learning. *Measurement: Sensors*, 2022. 2

[29] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv:2204.12484*, 2022. 2, 3, 4

[30] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *ICRA*, 2019. 2

[31] Ian T Young and Lucas J Van Vliet. Recursive implementation of the gaussian filter. *Signal processing*, 1995. 2

[32] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *ECCV*, 2022. 2

[33] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv:2207.06400*, 2022. 1, 2

[34] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 2021. 2

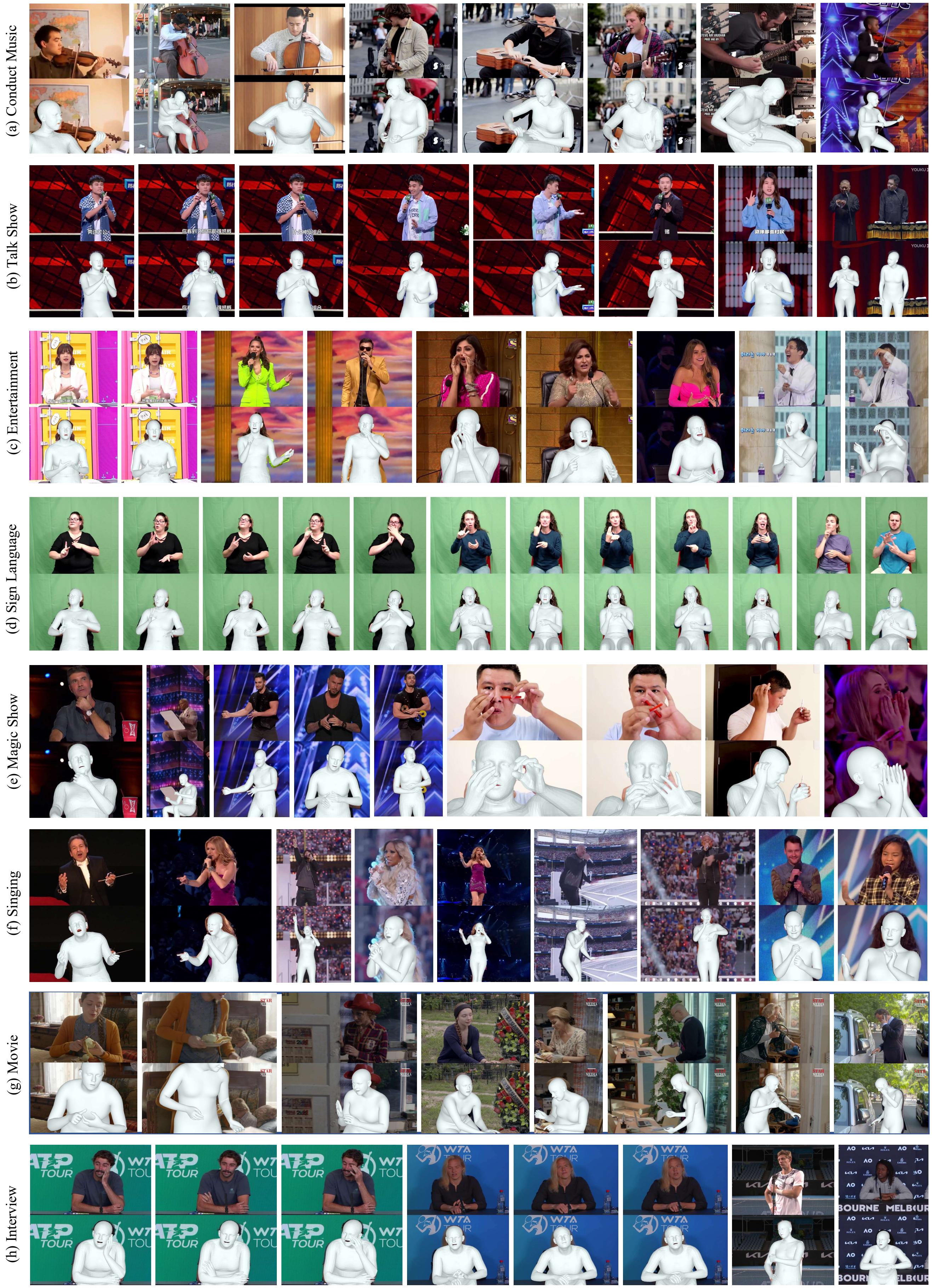


Figure S-5. Illustration of the ground-truth SMPL-X annotation for the eight scenes in *UBody*. For each scene, we show the input image (the upper) and our annotation (the lower).

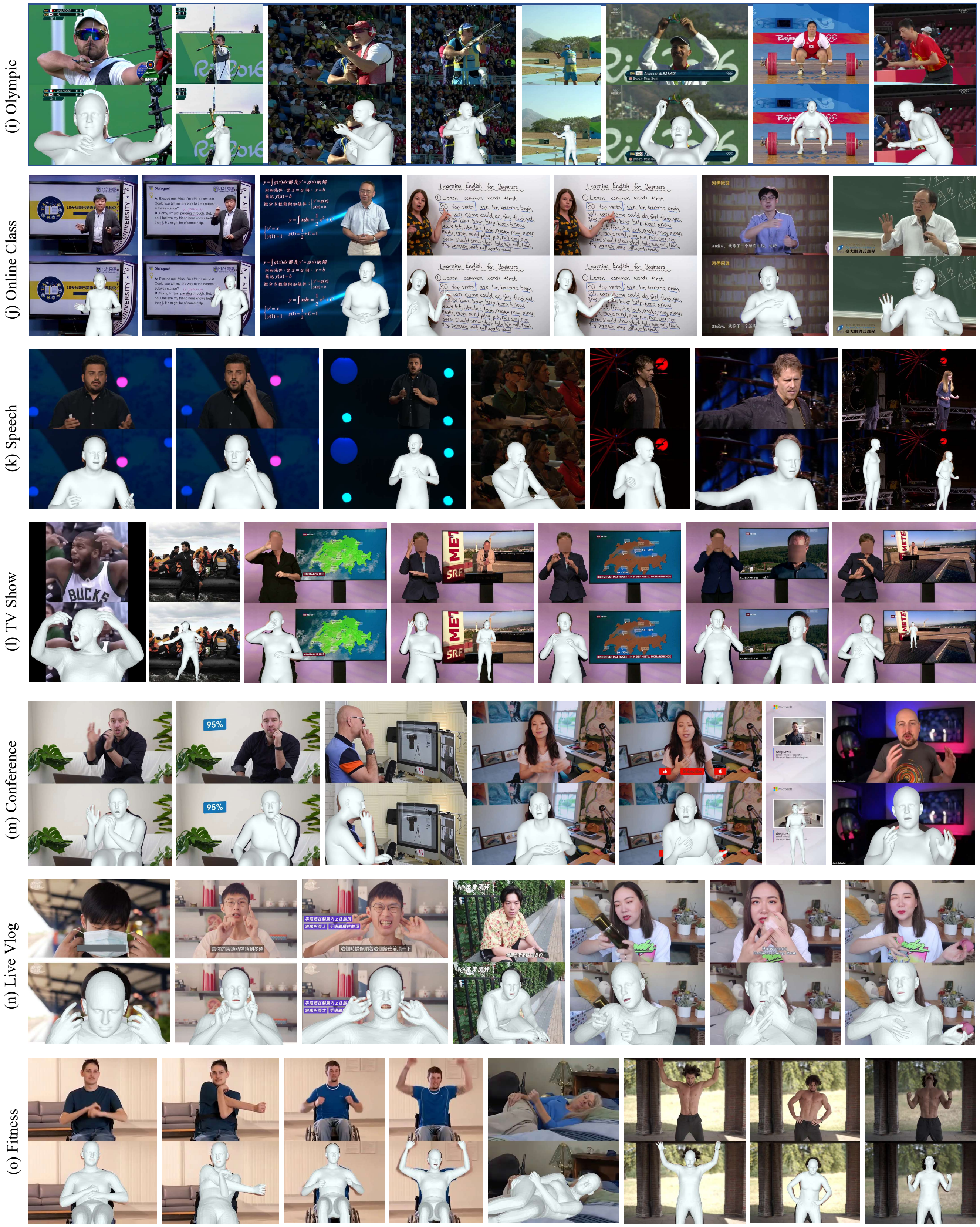


Figure S-6. Illustration of the ground-truth SMPL-X annotation for seven other scenes in *UBody*. For each scene, we show the input image (the upper) and our annotation (the lower).

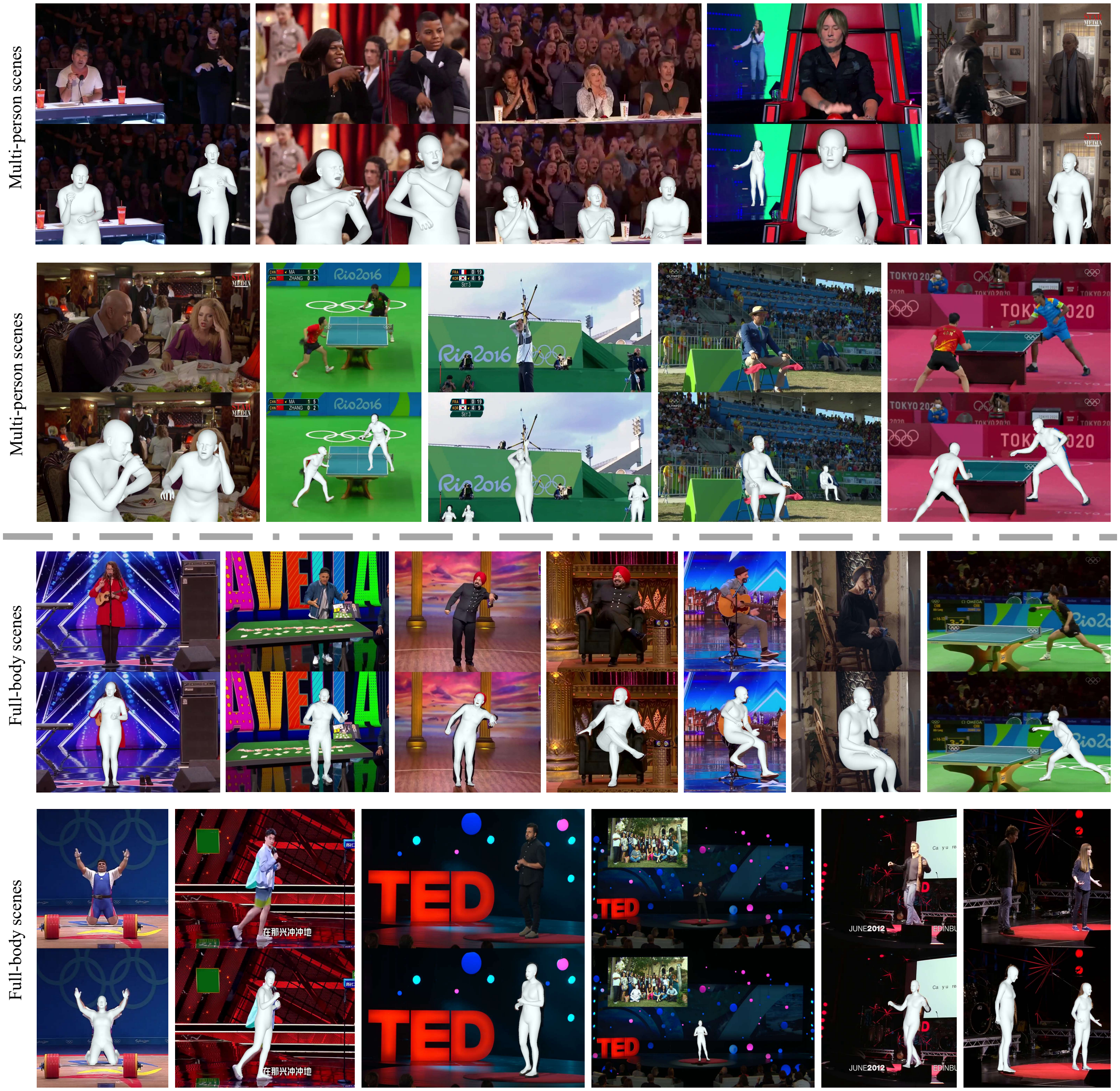


Figure S-7. Illustration of the ground-truth SMPL-X annotation for some special cases: *multi-person scenes* and *full body scenes* in *UBody*. Our annotation pipeline can still work well on these scenes.