AssemblyHands-X: Modeling 3D Hand-Body Coordination for Understanding Bimanual Human Activities

Tatsuro Banno   Takehiko Ohkawa   Ruicong Liu   Ryosuke Furuta   Yoichi Sato
The University of Tokyo, Tokyo, Japan
{banno, ohkawa-t, lruicong, furuta, ysato}@iis.u-tokyo.ac.jp

Abstract

*Bimanual human activities inherently involve coordinated movements of both hands and body. However, the impact of this coordination in activity understanding has not been systematically evaluated due to the lack of suitable datasets. Such evaluation demands kinematic-level annotations (*e.g.*, 3D pose) for the hands and body, yet existing 3D activity datasets typically annotate either hand or body pose. Another line of work employs marker-based motion capture to provide full-body pose, but the physical markers introduce visual artifacts, thereby limiting models' generalization to natural, markerless videos. To address these limitations, we present* AssemblyHands-X*, the first markerless 3D hand-body benchmark for bimanual activities, designed to study the effect of hand-body coordination for action recognition. We begin by constructing a pipeline for 3D pose annotation from synchronized multi-view videos. Our approach combines multi-view triangulation with SMPL-X mesh fitting, yielding reliable 3D registration of hands and upper body. We then validate different input representations (*e.g.*, video, hand pose, body pose, or hand-body pose) across recent action recognition models based on graph convolution or spatio-temporal attention. Our extensive experiments show that pose-based action inference is more efficient and accurate than video baselines. Moreover, joint modeling of hand and body cues improves action recognition over using hands or upper body alone, highlighting the importance of modeling interdependent hand-body dynamics for a holistic understanding of bimanual activities.*

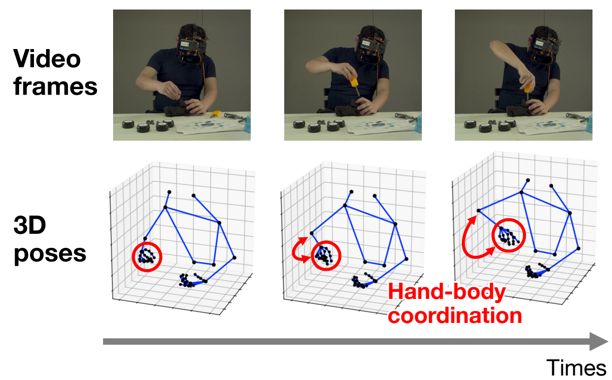


Figure 1. **Bimanual human activities inherently involve coordinated movements of both hands and body.** Our proposed benchmark, AssemblyHands-X, provides the first markerless 3D hand-body dataset, enabling systematic evaluation for the effect of hand-body coordination in action recognition.

1. Introduction

When we imagine nuanced everyday activities like screwing with a screwdriver, we intuitively grasp the deep coordination between our hands and body. As shown in Fig. 1, precise hand movements coincide with forearm rotation and elbow bends. This tight hand-body coupling forms the distinct motion patterns of bimanual activities. Observing this connection brings us a question: *how can modeling these interdependent hand-body dynamics improve the understanding of bimanual activities?*

However, answering the above question is difficult due to the lack of suitable benchmarks that provide kinematic-level annotations (*e.g.*, 3D pose) for hand-body dynamics. Most existing benchmarks for bimanual activities [16, 21] focus on video-based action recognition [10], lacking explicit 3D hand-body signals. While recent activity datasets provide 3D body pose [1, 2], they struggle to obtain detailed hand information. Due to the complex articulation of hands, they typically omit annotating hand pose [2] or provide limited hand signals (*e.g.*, three keypoints in [1]). Conversely, prior works on fine-grained hand interactions capture rich 3D hand poses from egocentric [17] or multi-view [6, 13] setups, while they typically neglect capturing body pose due to their tailored camera setups for close-up hand areas. Furthermore, while recent marker-based motion capture can record precise 3D hand-body motion [5, 20], the use of physical markers introduces inevitable visual artifacts, resulting in models with limited generalization to

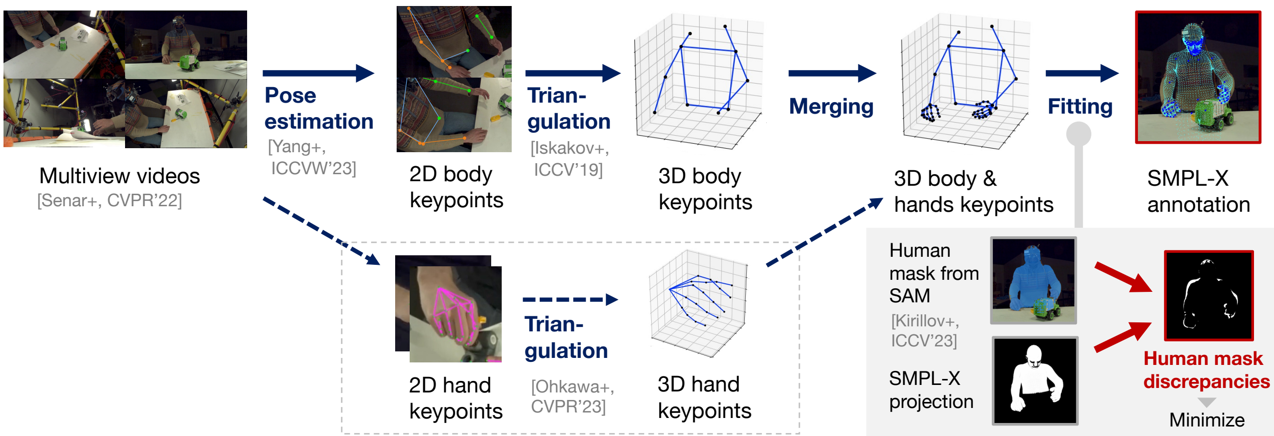


Figure 2. **Dataset annotation pipeline of AssemblyHands-X.** We first crop the body and hand regions, perform independent multi-view triangulation, and merge the results to obtain 3D hand-body keypoints. These keypoints are then used to fit the SMPL-X model, producing a unified 3D representation of human motion.

natural, markerless videos.

To address these limitations, we introduce *AssemblyHands-X*, the first markerless 3D hand-body benchmark for bimanual activities, designed for systematic evaluation for the effect of hand-body coordination on action recognition. Building on multi-view videos from Assembly101 [17], we develop an automatic annotation pipeline of 3D poses, which combines multi-view triangulation with SMPL-X [14] fitting. This enables the reliable recovery of synchronized 3D hand and body poses without relying on maker-based motion capture systems [5, 20].

Based on our new hand-body annotations, we validate different input representations for action recognition (*e.g.*, video, hand pose, body pose, or hand-body pose) on recent backbones using graph convolutional [12, 22] and spatio-temporal attention [18]. Our extensive experiments show that pose-based action inference is both more efficient and accurate than video-based baselines. This shows the effectiveness of using 3D pose as a compact and discriminative representation for action recognition. Furthermore, joint modeling of hands and body consistently improves action recognition compared to using either modality alone, highlighting that capturing interdependent hand-body dynamics benefits a holistic understanding of bimanual activities.

The main contributions of this work are as follows:

- We introduce AssemblyHands-X, the first markerless 3D hand-body benchmark for bimanual activities, designed for the systematic evaluation of the effect of hand-body coordination in action recognition.
- We introduce a multi-view annotation pipeline with triangulation and SMPL-X model fitting, which enables reliable capture of 3D poses for both hands and body without relying on maker-based motion capture systems.
- Using AssemblyHands-X, we evaluate different input representations for action recognition (*e.g.*, video, hand pose, body pose, or hand-body pose), demonstrating that modeling interdependent hand-body dynamics is crucial for a holistic understanding of bimanual activities.

2. Dataset annotation pipeline

2.1. Overview

An overview of our AssemblyHands-X dataset annotation pipeline is illustrated in Fig. 2. The input consists of multi-view videos from Assembly101 [17], and the output includes 3D keypoints for the hands and upper body, along with cooresponding parameters of a parametric whole-body human model (*i.e.*, SMPL-X [14]).

The pipeline begins by separately cropping the body and hands and performing multi-view triangulation independently for each. For hand pose estimation, we adopt the precise triangulation pipeline proposed by Ohkawa *et al.* [13]. In contrast, accurate body pose estimation during bimanual activities presents unique challenges, such as body truncation due to limited camera viewpoints. To address these issues, we propose a new body-specific triangulation pipeline (Sec. 2.2). We then combine the estimated 3D body and hand joints to generate coordinated 3D hand-body keypoint annotations. Finally, leveraging these 3D keypoints along with human segmentation masks extracted from the multi-view images [9], we fit the SMPL-X parametric model to each frame (Sec. 2.3), providing a kinematically consistent 3D representation of human motion.

The following sections provide a detailed description of each stage of the annotation pipeline, with particular emphasis on the body triangulation pipeline and the SMPL-X fitting process.

2.2. Body triangulation pipeline

2D keypoint detection. For each cropped frame, we first apply a 2D keypoint estimation model [19] for each camera view $c = 1, 2, \dots, C$, obtaining 2D locations $(x_c, y_c) \in \mathbb{R}^2$ and confidence scores $w_c \in [0, 1]$ for each body joint. We then apply median filtering to each keypoint across frames in every view to reduce short-term fluctuations.

Next, to suppress overconfident 2D keypoints near crop boundaries, we modulate each keypoint's confidence w_c based on its distance to the image edges. The adjusted confidence w'_c is computed as:

$$w'_c = \min\left(\frac{x_c}{m}, \frac{W - x_c}{m}, \frac{y_c}{m}, \frac{H - y_c}{m}, 1\right) \cdot w_c \quad (1)$$

Here, (x_c, y_c) is the keypoint location within a cropped size $W \times H$, and m is a predefined margin width. This formulation ensures that keypoints farther than m pixels from the crop edges retain their original confidence, while those within the margin are linearly downweighted as they approach the boundary. We then threshold the adjusted confidence scores at $\tau = 0.15$, setting all values below τ to zero and linearly rescaling values above τ to the range $[0, 1]$. This effectively discards low-confidence keypoints while preserving and normalizing reliable ones for subsequent processing.

3D keypoint triangulation. Next, we estimate 3D body joint locations by combining 2D keypoints from multiple views using a confidence-weighted linear triangulation [8]. Let the 2D position of a joint in camera $c = 1, \dots, C$ be $(x_c, y_c) \in \mathbb{R}^2$, with adjusted confidence score $w'_c \in [0, 1]$. Following [8], we construct a matrix $A_j \in \mathbb{R}^{2C \times 4}$ from the component of camera projection matrices of each view and 2D joint locations (see [7] for full details), and solve for the homogeneous 3D coordinate $\boldsymbol{X}_j \in \mathbb{R}^4$ using the following confidence-weighted least-squares equation:

$$(\boldsymbol{w}' \circ A)\boldsymbol{X} = 0 \quad (2)$$

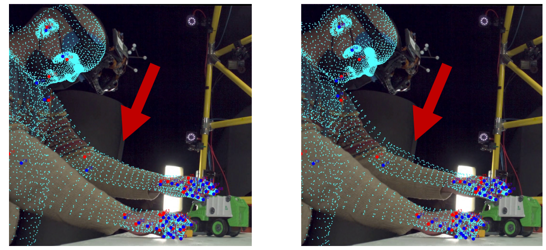
Here, $\boldsymbol{w}' = (w'_1, w'_1, w'_2, w'_2, \cdots, w'_C, w'_C)^\top$, and $\circ$ denotes the Hadamard product, meaning that the i-th row of matrix A is multiplied by the i-th element of the vector $\boldsymbol{w}'$.

This process yields the 3D locations of the body joints. By combining these with the precise 3D hand joint annotations from Ohkawa *et al*. [13], we generate a unified set of 3D keypoints covering the entire upper body and hands.

2.3. SMPL-X model fitting

Next, given the estimated 3D joint positions for both the hands and body, we optimize the SMPL-X parameters, namely the shape $\boldsymbol{\beta}$ and the pose $\boldsymbol{\theta}$, for each frame by minimizing the following objective:

$$\mathcal{L} = \lambda_{\text{joint}}\mathcal{L}_{\text{joint}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_\beta\|\boldsymbol{\beta}\|_2^2 + \lambda_\theta\|\boldsymbol{\theta}\|_2^2, \quad (3)$$



w/ $\mathcal{L}_{\text{mask}}$ w/o $\mathcal{L}_{\text{mask}}$

Figure 3. **Qualitative results of the fitting process with and without the silhouette loss $\mathcal{L}_{\text{mask}}$.** Incorporating $\mathcal{L}_{\text{mask}}$ significantly improves the alignment between the projected SMPL-X mesh and the observed human silhouette.

where each term is weighted by its corresponding coefficient λ. The first term, $\mathcal{L}_{\text{joint}}$, penalizes the L_2 distance between predicted and target 3D joint positions. The second term, $\mathcal{L}_{\text{mask}}$, is a silhouette loss [11] that aligns the SMPL-X mesh projection with human segmentation masks extracted from the multi-view images [9], significantly improving the match between the projected mesh and the observed human silhouette (see Fig. 3).

3. Experiments on action recognition

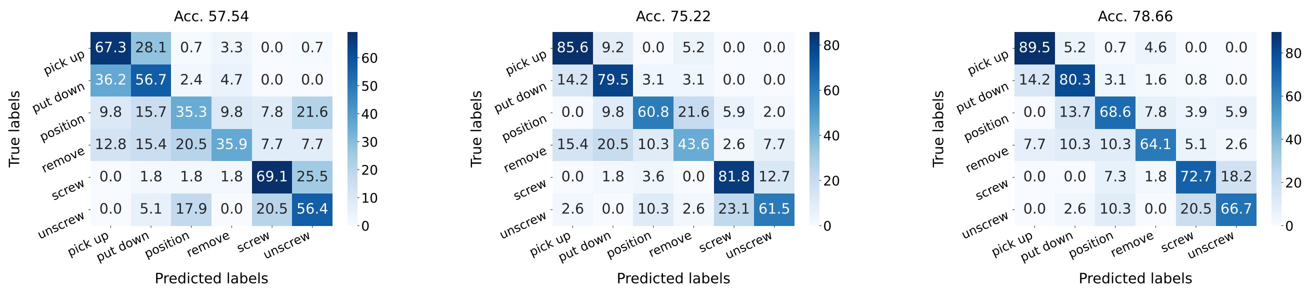
3.1. Experiment settings

Using AssemblyHands-X, we conduct a comprehensive benchmark of action recognition. As input, we utilize 3D keypoints obtained prior to SMPL-X fitting and the corresponding video frames from *View 3* of the Assembly101 [17], which provides a clear view of both the body and hands. For the action labels, we adopt the six verb classes defined in Assembly101, following prior work on 3D hand pose-based action recognition [13]: *pick up*, *put down*, *position*, *remove*, *screw*, and *unscrew*. All pose sequences are standardized to a fixed length of $T = 100$ frames, with shorter sequences padded by repeating the sequence.

3.2. Effect of input representations

We first evaluate different types of inputs on action recognition, including video, hand pose, body pose, and combined hand-body pose. For video-based inference, we consider two baselines: a large multimodal model (Gemini 2.5 pro [4]) with strong zero-shot capability of video comprehension and a standard video recognition model trained on our data (TSM [10]). For pose-based inference, we train graph convolutional networks (MS-G3D [12], BlockGCN [22]) and a transformer baseline with spatio-temporal attention (FreqMixFormer [18]).

Results. Table 1 shows action recognition result under different input representations. When comparing the best video-based model (TSM) and the best pose-based model



(a) Body pose (b) Hand pose (c) Combined hand-body pose

Figure 4. **Confusion matrices for action recognition under different 3D pose input configurations.** Each matrix corresponds to pose-based inference using FreqMixFormer [18], trained and evaluated on the respective 3D pose configuration.

Table 1. **Comparison of action recognition result under different input representations.** We report GFLOPs and Accuracy (%) of each model on the AssemblyHands-X benchmark. The dagger (†) indicates results obtained in a zero-shot evaluation setting.

	3D pose				
Method	Body	Hand	Video	GFLOPs	Acc.
Random choice	-	-	-	-	16.7
Gemini 2.5 Pro† [4]	✗	✗	✓	-	30.9
TSM [10]	✗	✗	✓	32.88	62.3
MS-G3D [12]	✓	✗	✗	0.79	53.0
	✗	✓	✗	8.24	63.4
	✓	✓	✗	10.16	69.6
BlockGCN [22]	✓	✗	✗	0.34	54.1
	✗	✓	✗	2.89	72.6
	✓	✓	✗	3.23	73.9
FreqMixFormer [18]	✓	✗	✗	0.46	57.5
	✗	✓	✗	3.63	<u>75.2</u>
	✓	✓	✗	4.03	**78.7**

(FreqMixFormer with 3D body-hand pose), the pose-based model achieves 16.4% higher accuracy while requiring 8.2× fewer GFLOPs. This demonstrates that 3D pose, which explicitly captures fine-grained hand-body kinematics, provides a compact and discriminative representation for recognizing actions in bimanual activities.

Table 1 also shows that, across all pose-based models, the full body-hand representation consistently outperforms either pose alone. This suggests that integrating full body-hand information contributes to more accurate recognition of bimanual activities. The corresponding confusion matrices for each 3D pose configuration (Fig. 4) further illustrate this effect, particularly showing that the hand-body model achieves substantial improvements on actions that are challenging for the hand-only model (*e.g.*, *position* and *remove*).

3.3. Effect of 3D pose annotation quality

Next, to investigate the impact of 3D pose annotation quality on action recognition, we evaluate both pose estimation using the latest single-view estimators and action recognition for varying inputs from these different pose estimators. For single-view estimation, we use SMPLer-X [3], a state-of-the-art 3D hand-body estimation model. Since SMPLer-X may produce noisy hand predictions, we also evaluate a variant where its hand estimates are replaced with those from HaMeR [15], a leading 3D hand estimator.

Table 2. **Comparison of action recognition result under different 3D annotation quality.** We report PA-MPJPE (mm) of the estimated 3D hand-body poses compared with our multi-view annotation data, alongside action recognition accuracy (%) on the AssemblyHands-X benchmark using MS-G3D [12].

		PA-MPJPE (mm)		
Annotation	#view	All	Hand	Acc.
SMPLer-X [3]	1	43.78	15.43	59.7
+ HaMeR [15]	1	45.47	10.62	<u>61.2</u>
Ours	8	-	-	**69.6**

Results. Table 2 reports the PA-MPJPE of the estimated 3D hand-body poses relative to our multi-view annotations, along with the resulting accuracy of a 3D pose-based action recognition model [12]. Our multi-view annotations significantly outperform all single-view baselines (SMPLer-X and SMPLer-X + HaMeR) in action recognition accuracy, demonstrating that higher-quality 3D poses contribute to better action recognition performance.

4. Conclusion

We introduce AssemblyHands-X, the first markerless 3D hand-body benchmark for bimanual activities, designed for the systematic evaluation of the effect of hand-body coordination in action recognition. Through extensive benchmarking, we demonstrate that modeling interdependent hand-body dynamics enhances a holistic understanding of bimanual activities.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Numbers JP23H00488 and JP24K02956.

References

[1] Dustin Aganian, Benedict Stephan, Markus Eisenbach, Corinna Stretz, and Horst-Michael Gross. Attach dataset: Annotated two-handed assembly actions for human action understanding. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 11367–11373, 2023. 1

[2] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 847–859, 2021. 1

[3] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 11454–11468, 2023. 4

[4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3, 4

[5] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12954, 2023. 1, 2

[6] Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, et al. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 428–448, 2024. 1

[7] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[8] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7718–7727, 2019. 3

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2, 3

[10] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7083–7093, 2019. 1, 3, 4

[11] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7708–7717, 2019. 3

[12] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 143–152, 2020. 2, 3, 4

[13] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023. 1, 2, 3

[14] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[15] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836, 2024. 4

[16] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11184–11193, 2021. 1

[17] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21096–21106, 2022. 1, 2, 3

[18] Wenhan Wu, Ce Zheng, Zihao Yang, Chen Chen, Srijan Das, and Aidong Lu. Frequency guidance matters: Skeletal action recognition by frequency-aware mixed transformer. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 4660–4669, 2024. 2, 3, 4

[19] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4210–4220, 2023. 3

[20] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 445–456, 2024. 1, 2

[21] Hao Zheng, Regina Lee, and Yuqian Lu. Ha-vid: A human assembly video dataset for comprehensive assembly knowledge understanding. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 67069–67081, 2023. 1

[22] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefine topology aware-

ness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2049–2058, 2024. 2, 3, 4