

Dyn-HaMR: Recovering 4D Interacting Hand Motion from a Dynamic Camera

Supplementary Material

The following sections supplement our main paper in terms of implementation details and additional qualitative and quantitative evaluations. Moreover, we refer the reader to our **supplementary videos** on the project webpage to better perceive the resulting motions and for a more comprehensive exposition, where we show promising results on in-the-wild hand reconstruction and qualitative comparisons against the state-of-the-art interacting hand reconstruction methods [13] under the challenging dynamic camera scenarios.

1. Implementation Details and Data Pre-Processing

We now provide more details into our initialization before moving onto the optimization scheme.

1.1. Initializing Motion States in Camera Frame

To initialize 2D observations in image plane and the MANO parameters ${}^c\mathbf{q}_t^h = \{\boldsymbol{\theta}_t^h, \boldsymbol{\beta}_t^h, {}^c\boldsymbol{\phi}_t^h, {}^c\boldsymbol{\tau}_t^h\}$ in the camera coordinate system at timestep t , we adopt a hierarchical pipeline.

First, we employ a 2D hand pose estimation model, ViTPose, following [13, 18], known for its performance in hand detection and palm localization. Despite its strength in detecting global hand regions, the model often produces jittery and inaccurate joint positions, making it insufficient for subsequent optimization processes. As a remedy, we refine the 2D inputs by cropping the image based on bounding boxes calculated from ViTPose’s 2D keypoint predictions. Specifically, for a set of keypoints ${}^{\text{vit}}\hat{\mathbf{J}}_t^h$, the bounding box is calculated by the point sets with a confidence filter $\epsilon_b = 0.5$ and an extension coefficient of 200%. To initialize the MANO parameters $\{\boldsymbol{\theta}_t^h, \boldsymbol{\beta}_t^h, {}^c\boldsymbol{\phi}_t^h\}$, we run the state-of-the-art hand reconstruction method of [13], using the officially released checkpoints, on the image patches restricted to the calculated bounding box. To estimate translation ${}^c\boldsymbol{\tau}_t^h = (x, y, d)$ of the two hands in 3D space, where d is the direction along depth, we simulate various versions of ${}^c\boldsymbol{\tau}_t^h$ in the camera coordinate system based on the predicted weak-perspective camera parameters (s, t_x, t_y) , assuming a fixed focal length $f = 1000$ following [13] using $x = t_x, y = t_y, d = \frac{2f}{s \times s_1}$, where s_1 is the image size. Alternatively, the camera translation in the camera coordinate system can also be acquired by solving the PnP algorithm (*i.e.* RANSAC [3]) with the 3D keypoints ${}^c\mathbf{J}_t^h$ and their corresponding 2D projections $\hat{\mathbf{J}}_t^h$ on the image plane. Finally, we infill the missed frames if the interval is less than 50 frames using the approach described in Sec. 3.1.

Keypoint refinement. Building upon the complete 3D ini-

tialization, we refine the corresponding 2D observations to improve accuracy and consistency. Specifically, we first detect all hands in the scene using ViTPose [18], and then combine these detections with predictions from MediaPipe [10] and the 2D re-projections derived from the 3D initialization. To achieve this, we extract wrist positions from ViTPose and pair them with finger joint predictions from MediaPipe, ensuring that both hands in each pair are correctly matched to the same individual. We replace the 2D finger joints whose confidence scores are lower than the threshold $\epsilon_j = 0.5$ with the corresponding 2D re-projections from the 3D initialization.

Handling occlusions. Modelling accurate interactions between hands is particularly challenging due to frequent occlusions, rapid motions, and truncations. These factors often lead to missed detections, especially in complex interaction scenarios. To address this, we employ a generative motion infilling approach, as detailed in Sec. 3.1. Specifically, we infill the hands from the timestep where it first appears to the last appearance timestep $(t_{\text{start}}, t_{\text{end}})$ with our generative motion prior. To handle the missed detections and occlusion more robustly, we only optimize the visible individual hands and mask out the objective terms for the occluded frames (*i.e.* we only update the latent code \mathbf{z} with these observed timesteps in Stage III), utilizing the motion prior as a guide to reason and infill the occluded frames.

Handling hallucinations. HaMeR can yield erroneous hallucinations – such as multiple hands in the same location, incorrect handedness, or implausible poses. Specifically, HaMeR’s detector can produce overlapping bounding boxes (bboxes) without suppression, leading to redundant or inconsistent predictions as both ViTPose and HaMeR process each bbox independently. As a remedy, we use ViTPose to extract 2D keypoints with confidences ($\text{IoU} > 0.9$) and instead of processing all overlapping detections, we retain only the bbox with the highest confidence before feeding it into HaMeR. We also filter any detection that appears in < 10 frames, reducing false positives. Incorrect handedness, where one hand is occasionally confused as its opposite, can be identified by the sudden change in the bbox (IoU with the previous bbox < 0.1), allowing us to mark these frames as invalid prior to generative infilling.

1.2. Optimization Scheme

Multi-stage optimization. Our key insight is to optimize the interacting hands in stages, balancing the per-frame motion accuracy and temporal smoothness while avoiding over-smoothing. We first optimize the two hands during

Table 1. **Acceleration analysis on HOT3D dataset.** Acc Err is reported w/o the div. of ω^2 (left) in mm/s^2 and with the div. of ω^2 in m/s^2 (right). Lower (\downarrow) is always better.

Method	Acc Err (w/o) \downarrow	Acc Err (with) \downarrow
ACR [20]	16.45	14.82
IntagHand [8]	15.12	13.62
HaMeR [21]	13.78	12.41
HaMeR + DPVO [17]	12.78	11.51
Ours (Dyn-HaMR)	4.95	4.46

Stage II individually with a lower $\lambda_{smooth} = 1$ to ensure accurate local pose and pixel alignment. After obtaining plausible global motion, we start to jointly optimize two hands in a single batch with interacting hand motion prior module, which makes the scale information shared between the two hands and further constrains the hands-camera displacement plausible. During optimization, the dimension of the latent code \mathbf{z} is 128 in the hand motion prior module. Interpenetration is only applied when both hands are present in the scene.

Chunk optimization. (i) For the pre-processing of long sequence $\mathcal{V} = \{\mathbf{I}_1, \dots, \mathbf{I}_T\}$ with $T \geq 128$ frames, we segment each video into chunks of 128 frames. This ensures compatibility with the hand motion prior module, which adopts a sequence length of 128 for motion parameterization as per [2]. Subsequently, we optimize the motion segments in chunks. We initialize the next motion and camera state with the end state of the last chunk (e.g. initialize \mathbf{C}_{127} and \mathbf{q}_{128}^h with the optimized output of \mathbf{q}_{127}^h and \mathbf{C}_{128}), as well as the world scale factor ω . (ii) In terms of the post-processing for evaluation, we align the translation parameters across segments and combine them to generate seamless visualizations of the reconstructed motion.

2. Evaluation Metrics

To compute **G-MPJPE** and **GA-MPJPE**, we first align the first two frames (G-MPJPE) or the whole sequence (GA-MPJPE) of hand motion with the GT using Umeyama method. We then transform the prediction to align with the GT before computing the MPJPE as the mean L_2 distance between each predicted and GT joint. To compute the **Acc Err**, we followed the common practice in HMP, GLAMR and SLAHMR ignoring the division, $\alpha_i^t = v_i^{t-1} - 2v_i^t + v_i^{t+1}$, where α_i^t, v_i^{t-1} are the acc. and velocity at timestep t without the division of discretized time step. We further analyze the effect of the different computation of acceleration where the discretized time step is applied or not. We report both Acc Err with and w/o division (mm^2 and m/s^2) in Tab. 1, where we can observe that our method keeps consistently better results under different computation method.

For the **RTE** (%) of sequence with N frames, we compute it the as the Trajectory Error as in Sec. 3:

$$\frac{1}{N} \sum_{i=1}^N \|\mathbf{T}_{\text{target}}^i - (\mathbf{R} \cdot \mathbf{T}_{\text{pred}}^i + \mathbf{t})\|_2 / \Delta, \text{ where } \Delta = \sum_{i=1}^{N-1} \|\mathbf{T}_{\text{target}}^{i+1} - \mathbf{T}_{\text{target}}^i\|_2, \text{ with } (\mathbf{R}, \mathbf{t}) \text{ being the computed rigid transform and } \mathbf{T} \text{ the root translation.}$$

3. Additional Experiments

In this section, we provide experiments for our 4D global hand motion reconstructions from both in-the-wild videos and existing benchmarks (i.e. H2O [6], FPHA [4], HOI4D [9], EgoDexter [12]).

Evaluation metrics. We evaluate both the reconstruction quality and the plausibility of our motion. In addition to the evaluation metrics of (i) local hand pose and shape, (ii) global hand motion, we further conduct (iii) **motion plausibility evaluation** quantifying the plausibility and fidelity of our bimanual reconstructions. In addition to the metrics introduced in our main paper such as MPJPE (mm), PA-MPJPE (mm) and Acc Err (mm/s^2), we further propose the following two fronts for evaluation:

- **Global trajectory plausibility:** We quantify Trajectory Error (Trans Err) in % for each clip after the rigid alignment and normalize it by displacements of ground truth trajectory.
- **Bimanual (interacting hand) pose plausibility:** We report Fréchet Distance (FID) between estimations and the GT data to quantify the plausibility of the joint pose of interacting hands. To evaluate the smoothness and the interaction quality, we compute Jerk ($10m/s^3$) in the world coordinate system and Mean Inter-penetration Volume (Pen) in cm^3 to measure between the two hands.

Specifically, we adopt a PointNet++ [14] based embedding network for Fréchet distance on latent space following [7, 16]. For single-hand plausibility, we train it on the combined dataset of InterHand2.6M and H2O to regress the local hand poses $\theta \in \mathbb{R}^{3 \times 15}$ in axis-angle representation from the hand mesh vertices $\mathbf{V} \in \mathbb{R}^{3 \times 778}$. We keep the original setting while only modifying the last layer. The reconstruction MPVPE achieves $1.18mm$. For interacting hands, we modify the last layer to regress the hand pose for both hands as well as the relative root translation and rotation. This achieves $1.65mm$ and $1.61mm$ MPVEP for the left and right hands, respectively. We report the FID score for both the single hand version and the bimanual version, whenever two hands are jointly visible.

3.1. Results

Local motion estimation. To fully analyze the effectiveness of our pipeline, we conduct experiments on **FPHA** [4], an egocentric RGB-D hand-object motion dataset, which contains 105K frames encompassing 45 daily hand action categories, captured across diverse hand configurations with ground truth 3D hand joint annotations provided in the camera coordinate system. We evaluate the quality of biman-

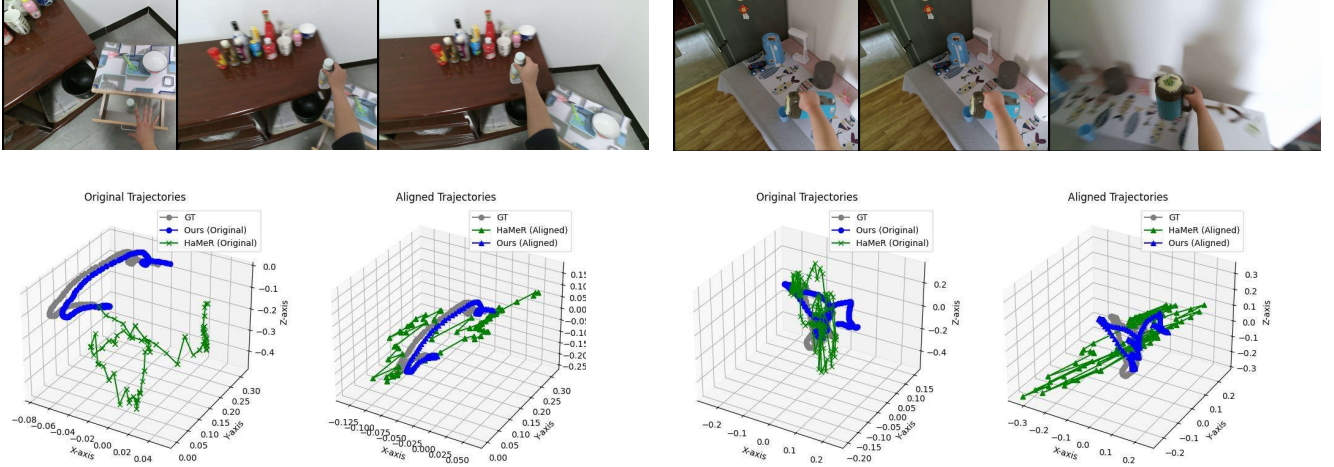


Figure 1. Comparison of global trajectory on HOI4D [9].

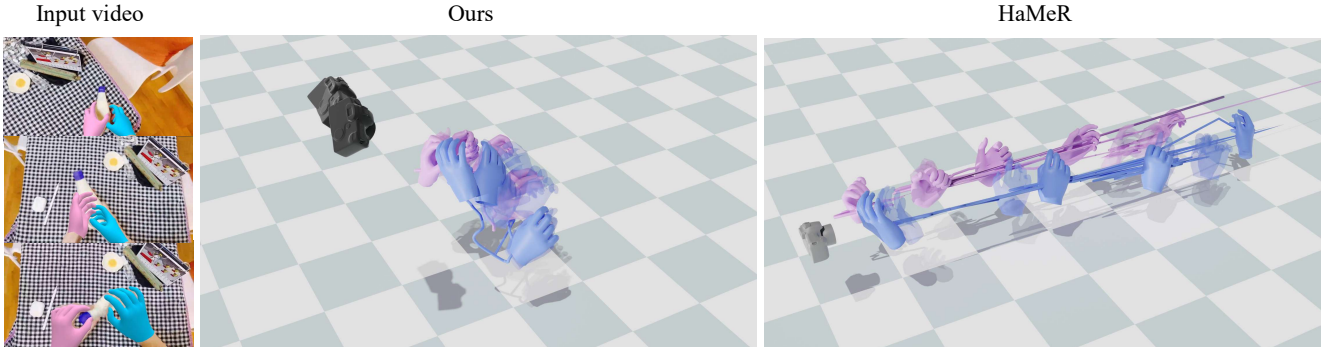


Figure 2. Qualitative comparison with state-of-the-art method HaMeR [21] on in-the-wild online videos.

ual hand reconstruction by measuring the root-relative joint reconstruction error (local motion) using metrics such as MPJPE, PA-MPJPE, and Acc Err. As shown in Tab. 2, our method demonstrates superior performance, achieving the lowest MPJPE (18.9 mm) and PA-MPJPE (12.5 mm) compared to other state-of-the-art approaches. Additionally, we achieve a competitive Acc Err of **5.7**, demonstrating improved temporal smoothness and consistency. Furthermore, qualitative results in Fig. 3 further illustrate the robustness of our pipeline in handling complex in-the-wild scenarios such as egocentric hand-object interactions and hand-hand interactions.

Global motion estimation. As introduced in Sec. 4.1 of the main paper, H2O [6] and HOI4D [9] contain dynamic camera videos with available camera poses to convert the hand poses from the camera coordinate system to the world coordinate system. To evaluate global motion recovery performance, we have conducted qualitative evaluations on the aforementioned four egocentric hand-object interaction datasets mentioned in Sec. 4.1 as well as on in-the-wild videos. In this section, we first present qualitative results on

these datasets shown in Figs. 5, 7 and 8, where our method produces plausible 4D global motion with trajectories in the world coordinate system, while previous state-of-the-art methods fail to capture the global motion in 3D space, especially from dynamic cameras. Furthermore, our method yields more plausible depth reasoning in the bimanual setting and significantly reduces the jitter in translations. To quantify the reconstruction accuracy and errors, we evaluate the Translation Error and Jerk in Tab. 3, where we can observe significant improvements over the state-of-the-art approaches. Specifically, we evaluate the RTE score for each of the motion trajectories in the world coordinate system. It can be observed that our method consistently archives the lowest translation error across datasets. We also conduct further analysis on the HOI4D dataset as shown in Fig. 1, which contains large camera displacements. Notably, we achieve 3.89% in Trans Err against 18.98% of HaMeR [13] on this dataset. We encourage to browse the reconstruction results in our **supplementary videos**, which clearly demonstrates the superiority of our pipeline against the state-of-the-art methods.



Figure 3. **Qualitative results of hand motion estimation under complex hand interactions and hand-object interaction.** In the figure, (a)-(b) show the samples from InterHand2.6M dataset [11], while (f)-(i) are from H2O [6], FPFA [4], HOI4D [9] and EgoDexter [12], respectively. Finally, (e) and (j) are reconstruction results from in-the-wild web videos.

Table 2. **Quantitative comparison on FPFA [4] dataset.** PA-MPJPE represents the MPJPE after Procrustes Alignment.

Method	MPJPE ↓	PA-MPJPE ↓	Acc Err ↓
ACR [20]	43.6	35.1	13.1
IntagHand [8]	41.2	31.6	12.4
HaMeR [13]	29.9	18.7	12.5
w/o bio. const.	19.6	13.5	6.1
w/o pen. const.	21.3	15.7	5.4
Ours (Dyn-HaMR)	18.9	12.5	5.7

Bimanual hand pose plausibility. In addition to evaluating global motion recovery, we conduct extensive experiments on complex interacting hand scenarios and assess the plausibility of the results. Fig. 4 indicates significantly more details in reconstruction in favor of our method, more stable **depth reasoning**, and higher local hand pose accuracy under self-occlusions compared to the baseline [13]. We further provide a plausibility evaluation of the 4D motion reconstructions in Tab. 3, where our approach is compared against state-of-the-art methods [8, 13, 20] and ablations of different modules. Our method consistently outperforms existing approaches by a large margin across all reported metrics. Thanks to the integration of penetration and biome-

chanical constraints, our approach exhibits superior stability in recovering bimanual poses from complex interacting scenarios, achieving the lowest FID and Jerk. This demonstrates the effectiveness of our Stage III in addressing the challenges of bimanual interactions.

3.2. Analysis

Ablation study of initialization. While our network is not restricted to any specific initialization backbone we provide an additional ablation study on the 3D motion state initialization to fully assess the effect of each component, where we conduct experiments on ACR [20], IntagHand [8] and HaMeR [13]. As illustrated in Tab. 4, we compare our full pipeline (HaMeR [13] initialization) with initializations from [8, 20] and Ours (Base) represents to initialize from the default MANO mean pose. It can be observed that there is a boost in the performance with the recent large-scale model based hand reconstruction framework HaMeR [13], which indicates that a better initialization could further improve the optimization and speed up the convergence.

Runtime. Our network is agnostic to the initialization method (e.g. camera, hand initialization), which affects the processing time. As shown in Tab. 5. On an NVIDIA A100

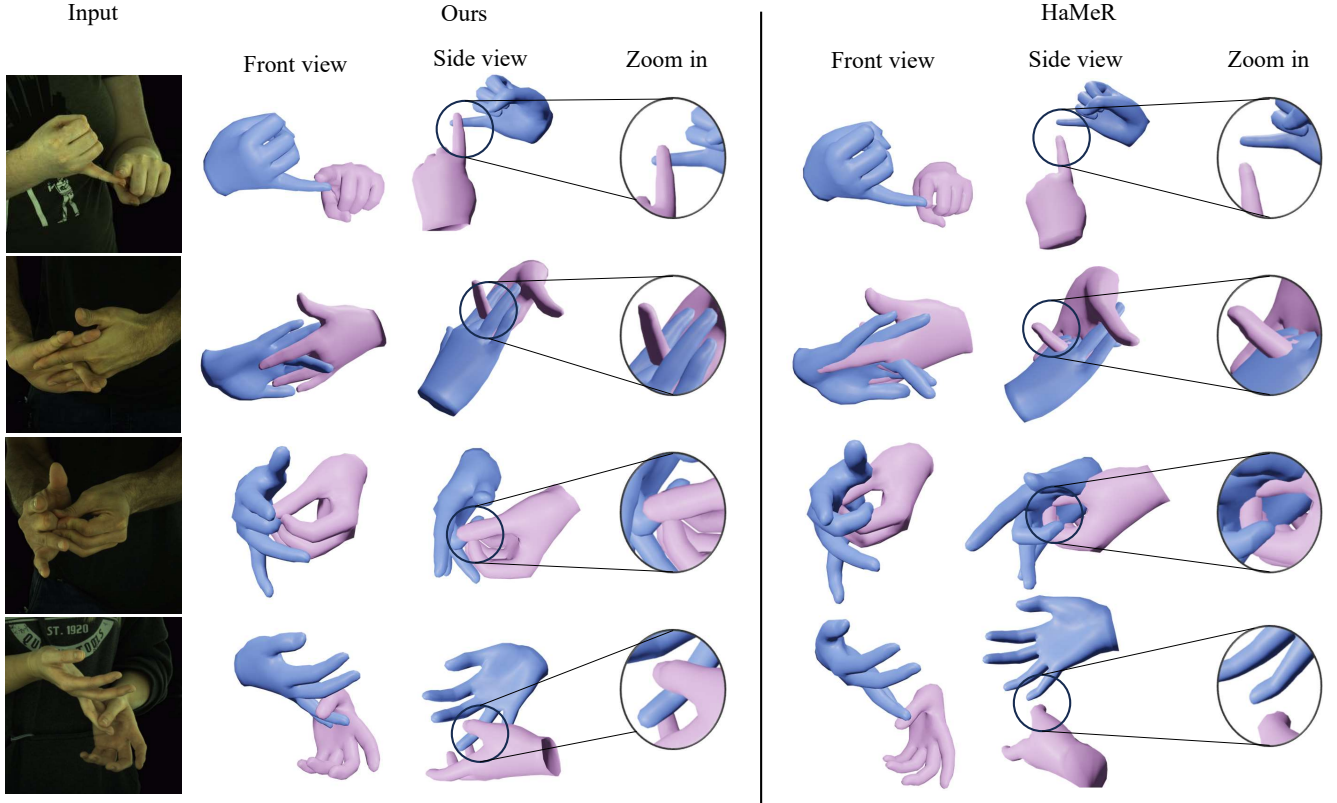


Figure 4. **Comparison with state-of-the-art hand reconstruction approach (static camera) on InterHand2.6M dataset [11].** We compare our method with state-of-the-art hand reconstruction approach HaMeR [21] under challenging hand interactions.

Table 3. **Plausibility evaluation on multiple datasets.** Results are reported on the H2O [6] and InterHand2.6M [11] to analyze the jitter, penetration, translation, and plausibility. FID is reported for both single hand (left) and two hands (right).

Method	H2O				InterHand2.6M			
	Jerk ↓	Pen ↓	Trans Err ↓	FID ↓	Jerk ↓	Pen ↓	Trans Err ↓	FID ↓
ACR [20]	149.43	0.07	10.89	1.95 / 4.45	153.62	5.05	8.65	2.51 / 5.36
IntagHand [8]	166.38	0.06	11.15	2.14 / 4.12	165.31	4.82	9.19	2.69 / 5.07
HaMeR [21]	195.77	0.06	10.43	1.76 / 4.78	183.45	5.17	8.43	2.45 / 5.45
Ours (w/o bio. const.)	2.65	0.04	4.71	1.89 / 2.78	4.57	2.67	4.41	1.89 / 4.12
Ours (w/o pen. const.)	2.36	0.02	4.13	1.38 / 2.12	4.03	4.23	4.93	1.53 / 4.64
Ours (w/o III)	2.98	0.02	4.21	2.01 / 2.93	4.81	4.49	4.96	2.89 / 4.87
Ours (Dyn-HaMR)	2.34	0.009	5.67	1.34 / 1.98	4.26	2.46	4.35	1.49 / 3.56

Table 4. **Ablation study on H2O [6] dataset.** To quantify the importance of the initialization, we compare the performance initialized from different state-of-the-art approaches [8, 13, 20].

Method	G-MPJPE ↓	GA-MPJPE ↓	MPJPE ↓	Acc Err ↓
Ours (Base)	55.8	47.6	28.9	4.2
Ours (ACR [20])	49.8	37.3	23.2	4.7
Ours (IntagHand [8])	48.9	41.4	25.1	4.5
Ours (HaMeR [21])	45.6	34.2	22.5	4.2
Ours (Long)	69.5	49.1	22.3	4.2

GPU, our experiments for a 128-frame video clip adopt HaMeR and ACR for 3D initialization, taking 3.18 minutes, and 8 seconds on average, respectively. Subsequently, optimizing stage II takes around 2.3 minutes. Finally, the

last stage takes 1 to 2 minutes. We use Pyrender for off-screen rendering, which takes an additional minute. Note, the rendering time can also vary depending on the specific resolution and number of views desired. Compared to existing optimization-based pipelines such as humor [15] and slahmr [19], which take more than 45 minutes to 2 hours for 128 frames on A100 GPU, our method achieves the fastest test time optimization, which makes a step towards efficient and real-time applications.

Long sequences degeneration. As described in Sec. 1 and 3, the errors in estimated global trajectories would accumulate over time in our moving camera setting. Therefore,

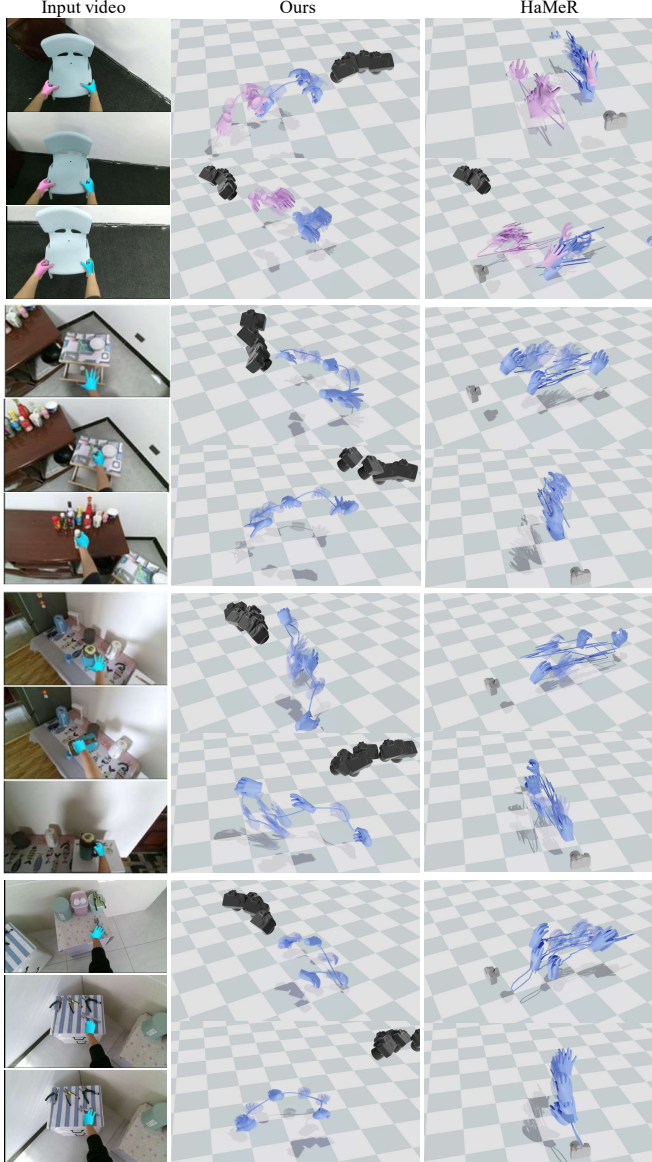


Figure 5. **In-the-wild 4D global hand motion reconstructions on HOI4D dataset [9].** We visualize the front view and the bird’s eye view, which are the upper row and the lower row in each of the sample motions. State-of-the-art hand reconstruction approach HaMeR [13] fails to recover plausible global trajectories while our method produces. Moreover, our method produces significantly less jitter and more plausible depth reasoning. Please see the [supplementary video](#) for better visualization of motions.

we follow standard evaluations for open-loop reconstruction (*e.g.*, SLAM and inertial odometry) to compute errors using a sliding window, similar to WAHM and GLAMR. To quantify its impact, we provide the results for *long* sequence here in Tab. 4, where we conduct the evaluation based on the original video sequence length instead of the 128 clips mentioned in the experiments section.

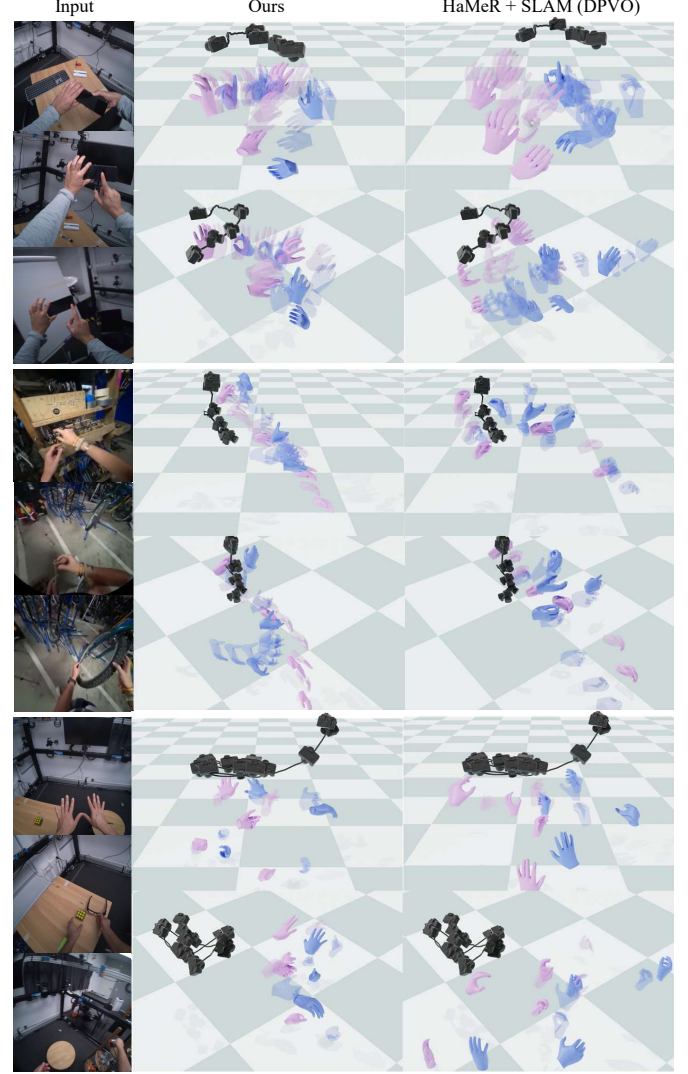


Figure 6. In-the-wild 4D global hand motion reconstructions on HOT3D [1] (top and bottom rows) and Ego-Exo4D [5] (middle row) datasets.

Table 5. **Runtime.** We show the individual runtime for initialization and each optimization stage separately.

Methods	Avg. runtime (min.)
HuMoR [15]	58.7
SLAHMR [19]	65.5
Camera tracking (DPVO [17])	1.49
3D Hand tracking (HaMeR [21])	3.18
3D Hand tracking (ACR [20])	0.13
2D keypoints detection [10, 18]	1.45
Stage II optimization	2.5
Stage III optimization	1.69
Dyn-HaMR (initialization)	3.07~6.12
Dyn-HaMR (optimization)	4.19

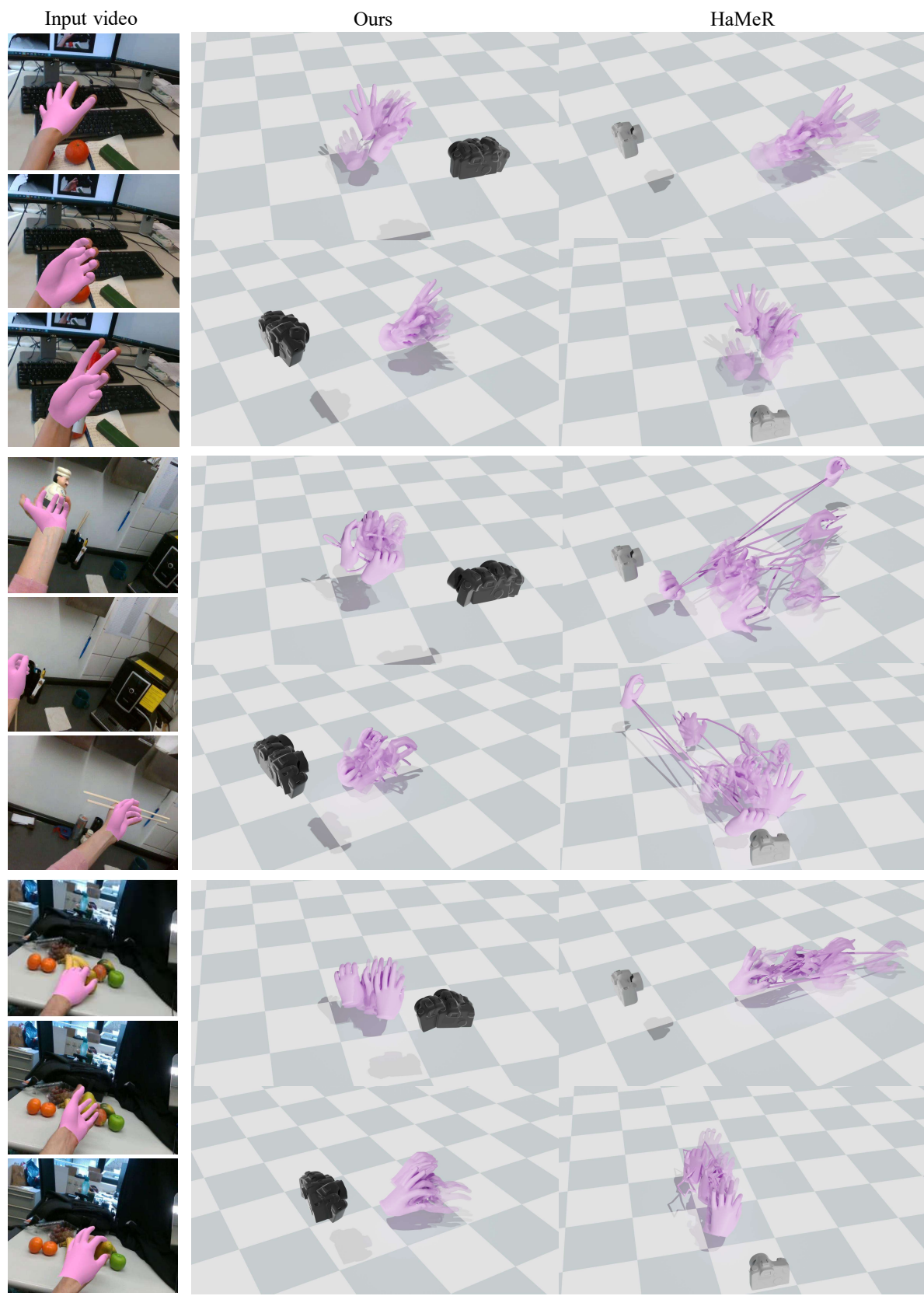


Figure 7. **In-the-wild 4D global hand motion reconstructions on EgoDexter dataset [12].** Please see the supplementary video for the motion visualization. Our method produces plausible global motion and depth reasoning.

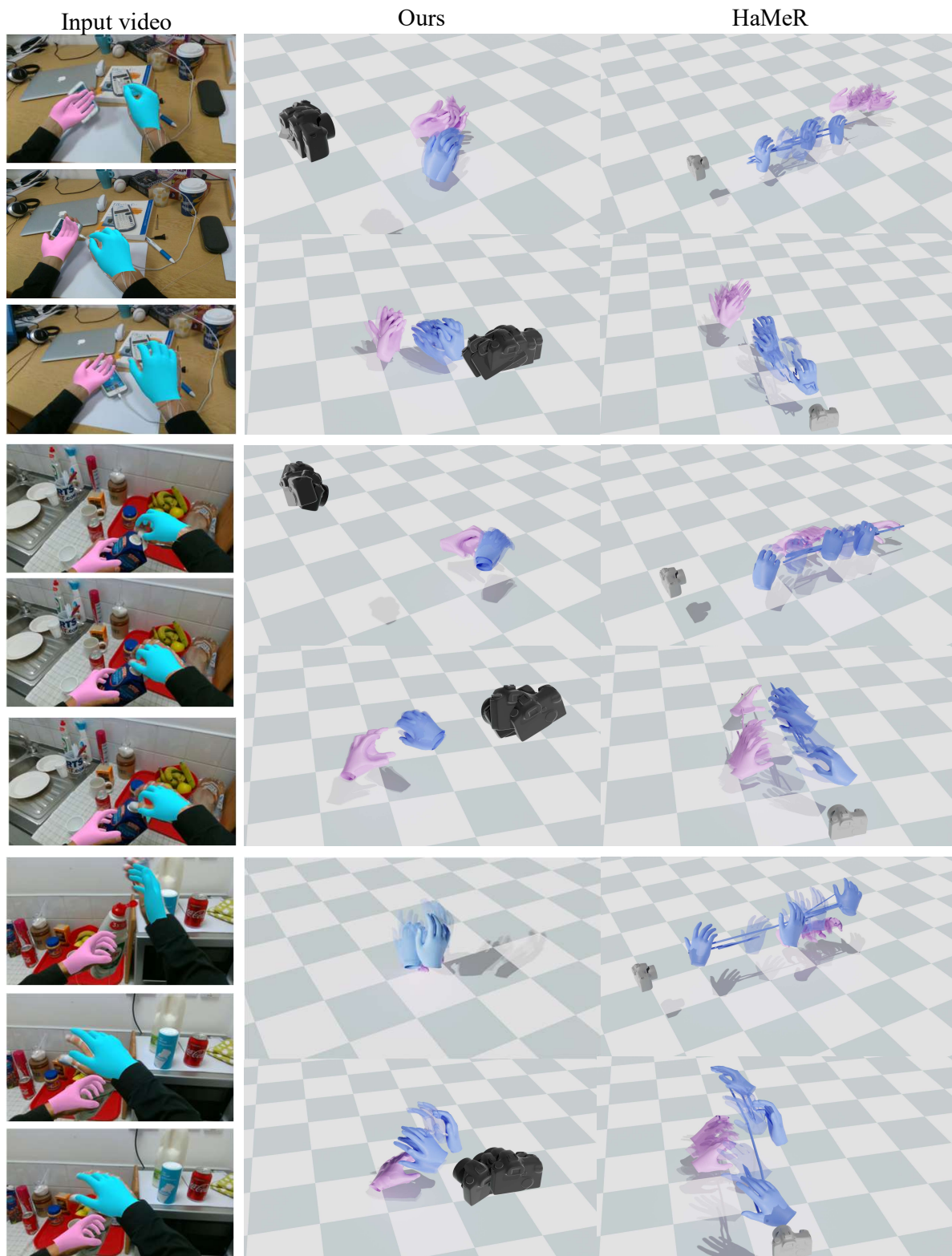


Figure 8. **In-the-wild 4D global hand motion reconstructions on FPHA dataset [4].** We also provide detailed motion visualization of FPHA in the supplementary video.

References

- [1] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, R. Newcombe, R. Wang, J. Engel, and T. Hodan. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. *arXiv preprint arXiv:2411.19167*, 2024. 6
- [2] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. Hmp: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2
- [3] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1
- [4] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4, 8
- [5] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 6
- [6] Taekwon Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021. 2, 3, 4, 5
- [7] Jihyun Lee, Shunsuke Saito, Giljoo Nam, Minhyuk Sung, and Tae-Kyun Kim. Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. In *CVPR*, 2024. 2
- [8] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4, 5
- [9] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 2, 3, 4, 6
- [10] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuoling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 1, 6
- [11] Gyeongsik Moon, Shou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 4, 5
- [12] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017. 2, 4, 7
- [13] 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*, pages 9826–9836, 2024. 1, 3, 4, 5, 6
- [14] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2
- [15] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 5, 6
- [16] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3859–3868, 2019. 2
- [17] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 6
- [18] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1, 6
- [19] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5, 6
- [20] Zhengdi Yu, Shaoli Huang, Fang Chen, Toby P. Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 5, 6
- [21] Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9054–9064, 2023. 2, 3, 5, 6