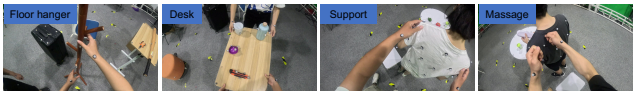
Perceiving and Acting in First-Person: A Dataset and Benchmark for Egocentric Human-Object-Human Interactions

Supplementary Material

A. Object Setting

We list all 50 adopted objects of the InterVLA dataset in Tab. A, which include 35 small objects and 15 large objects. Based on statistics, **41/100** scripts involves large object manipulations. For human-human interactions, InterVLA consists almost entirely of indirect human-human interactions through objects where the assistants need to comprehend the intention of the instructor before making responses. We also include some pure human-human interactions like "support someone" and "wave". We provide more examples of large object manipulation and pure human-human interactions in Fig. A for better illustration.



(a) Interactions with large objects (b) Pure human-human interactions

Figure A. **InterVLA samples.** More samples of large object manipulation and pure human-human interactions of InterVLA.

B. SMPL Optimization

Formally, the SMPL parameters consist of the body pose parameters $\theta \in \mathbb{R}^{N\times 23\times 3}$, root translation $\gamma \in \mathbb{R}^{N\times 3}$, global orientation $q \in \mathbb{R}^{N\times 3}$, and the shape parameters $\beta \in \mathbb{R}^{N\times 10}$, where N indicates the number of frames. We initialize the shape of the participant β based on their height and weight as [8]. Then, we optimize the SMPL parameters based on the Mocap data with the following optimization objective as:

$$\mathcal{L} = \lambda_j \mathcal{L}_j + \lambda_s \mathcal{L}_s + \lambda_{reg} \mathcal{L}_{reg}, \tag{1}$$

where

$$\mathcal{L}_j = \frac{1}{N} \sum_{i=0}^{N} \sum_{j\in\mathcal{J}} \| \boldsymbol{J}_j^i(\mathbb{M}(\theta,\gamma,q) - \boldsymbol{g}_j^i \|_2^2 \tag{2}$$

aims to fit the SMPL joints to our captured skeleton data, where $\mathcal{J}$ denotes the joint set, $\mathbb{M}$ is the SMPL parametric model, $\boldsymbol{J}_j^i$ is the joint regressor function for joint j at i-th frame, $\boldsymbol{g}$ is the Mocap skeleton data. A smoothing term

$$\mathcal{L}_s = \frac{1}{N-1} \sum_{i=0}^{N-1} \sum_{j\in\mathcal{J}} \| \boldsymbol{J}_j^{i+1} - \boldsymbol{J}_j^i \|_2^2 \tag{3}$$

is applied to alleviate the pose jittering between frames. A regularization term

$$\mathcal{L}_{reg} = \|\theta\|_2^2 \tag{4}$$

is applied to constrain the pose parameters from deviating.

C. Hand Pose Results

We highlight the dexterous hand gestures such as manipulating objects and interacting with other individuals. However, attaching additional reflective markers on the hands fails to yield robust finger gestures empirically, and employing heavy inertial gloves significantly compromises the fidelity of RGB videos. To this end, we prioritize the natural RGB data of hand interactions and attach only three reflective markers to the hands to determine the rotation of the wrists. We notice that existing hand pose estimation algorithms [3, 5–7, 14] demonstrate impressive accuracy and robustness even for in-the-wild hand images while other works [1, 2, 12, 13] jointly estimate the poses of both hands and the interacting objects. To this end, we apply the state-of-the-art hand pose estimation methods [7] on our head-mounted egocentric videos as shown in Fig. B, which yield robust estimated hand poses.

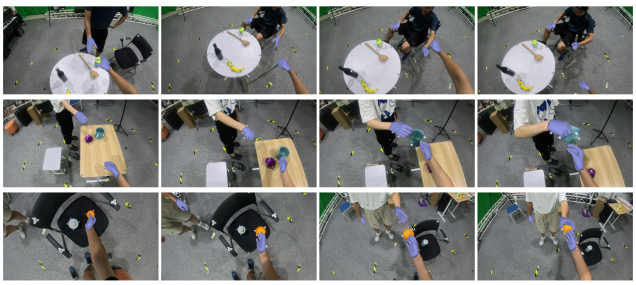


Figure B. **Hand Pose Reconstruction**. Visualization results of the hand pose estimation results performed by WiLoR [7] on the head-mounted egocentric videos of InterVLA.

We provide more visualization results of the failure cases of hand pose reconstruction of our InterVLA dataset by WiLoR [7] in Fig. C. We find that the state-of-the-art hand pose reconstruction method still fails to obtain smooth and accurate estimation results, which further validates the challenge of InterVLA. The failure parts are highlighted as red dashed boxes.

D. Continuity between commands

The continuity of operating one group of objects or one specific object is strictly guaranteed, such as "give me the bottle" → "pure some wine into the bottle" → "put the bottle back on the table". However, we don't emphasize the con-

01. Apple	02. Banana	03. Cucumber	04. Potato	05. Onion	06. Avocado	07. Orange
08. Liver bottle	09. Mid autumn box	10. Toothbrush box	11. Knife	12. Spatula	13. Ladle	14. Spoon
15. Fork	16. Football	17. Mouse	18. Slipper left	19. Slipper right	20. Remote control	21. Wine bottle
22. Wine bottle black	23. Wine cylinder	24. Beer bottle	25. Teapot	26. Tape	27. Mug 1	28. Mug 2
29. Vacuum cup	30. Hammer	31. Piler	32. Screwdriver	33. Utility knife	34. Fruit knife	35. Rubbish bin
36. Ukulele	**37. Big box 1**	**38. Big box 2**	**39. Suitcase**	**40. Baseball bat**	**41. Besom**	**42. Dustpan**
43. Floor hanger	**44. Camera mount**	**45. Chair 1**	**46. Chair 2**	**47. Chair square**	**48. Sofa chair**	**49. Desk**
50. Desk circle						

Table A. **The objects setting of the InterVLA dataset.** The first 35 items are small objects, while the remaining 15 are large objects highlighted in **bold** font.

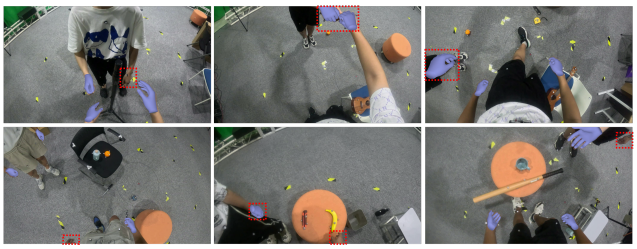


Figure C. **Failure Cases of Hand Pose Reconstruction**. We provide more results of the failure cases of the hand pose estimation results performed by WiLoR [7] on the head-mounted egocentric videos of InterVLA.

tinuity among different objects with different functionalities. After the temporal segmentation process, all the atomic commands serve as independent VLA segments with complete semantics.

E. Interaction Synthesis Settings

MDM [10]. We extend the original human motion generation model to human-object-human interaction generation, where the feature dimensions of the input and output are extended from D_h to $D_h + D_o$, where D_h is the dimension of human motion and D_o denotes that of object motion. To embed the condition input of object geometries, we feed them into a linear layer and concatenate them with the initial poses of the objects. Then, all the conditions are concatenated with the noised input into the motion embedding.

PriorMDM [9]. The original PriorMDM [9] is designed for two-person motion generation with two dual branches of MDM [10] and ComMDM to coordinate these two branches. We modify the two branches into a human motion branch and an object motion branch. Besides, we place the ComMDM module after the 4-th transformer layer of each branch to enable communication between the two branches.

HIMO [4]. HIMO was designed for single-person interaction with multiple objects. We extend this method to two persons and up to seven objects. The object features are all concatenated together with the initial poses of these objects as the condition.

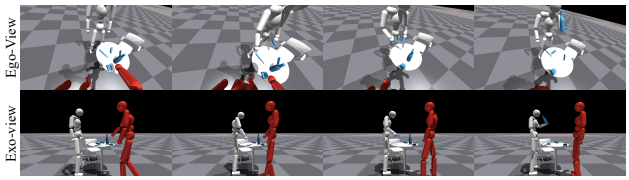


Figure D. InterVLA in simulation for embodied intelligence.

F. Experimentation on embodied intelligence.

We consider this an important direction and will continue to investigate in subsequent work. Currently, following your advice, we have transferred InterVLA to Isaac Gym as Fig. D, trained a humanoid VLA model via behavior cloning inspired by HumanVLA [11], and conducted closed-loop replay tests within the simulator. The tracking results show that over **93.0%** of frames achieve an average body position error within 0.3m, which validates the accuracy of the captured motions and the utility for embodied applications.

G. Limitations

While InterVLA is the first dataset designed for AI assistants where both the versatile human-centric interactions and egocentric perspective are considered, we highlight that some limitations remain. 1) First, InterVLA is limited to indoor scenarios with 50 daily objects involved. Extending our setting to outdoor settings or enriching the scenes are of great merit. Besides, indoor-captured dataset lack a certain level of realism, which is a common issue among indoor motion capture datasets. However, as the first dataset of its kind, we believe it holds significance for the broader human-robot interaction community. 2) Second, building InterVLA demands substantial time investment for attaching reflective markers, staging and changing the scenes and data processing. We strive to present InterVLA with >10 hours of high-quality interactive data, yet it is still insuf-

ficient for training large generalist interaction models. 3) Third, we discard the inertial gloves for capturing hand movements to preserve RGB realism. We apply several hand motion recovery models to InterVLA as illustrated before with extensive results and analysis.

References

[1] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 1

[2] 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 *European Conference on Computer Vision*, pages 428–448. Springer, 2025. 1

[3] Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 758–767, 2023. 1

[4] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, et al. Himo: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision*, pages 300–318. Springer, 2025. 2

[5] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1496–1505, 2022. 1

[6] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

[7] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. *arXiv preprint arXiv:2409.12259*, 2024. 1, 2

[8] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H Bülthoff, and Michael J Black. The virtual caliper: Rapid creation of metrically accurate avatars from 3d measurements. *IEEE transactions on visualization and computer graphics*, 25(5): 1887–1897, 2019. 1

[9] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2

[10] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2

[11] Xinyu Xu, Yizheng Zhang, Yong-Lu Li, Lei Han, and Cewu Lu. Humanvla: Towards vision-language directed object rearrangement by physical humanoid. *arXiv preprint arXiv:2406.19972*, 2024. 2

[12] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3895–3905, 2022. 1

[13] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19717–19728, 2023. 1

[14] Zhengdi Yu, Shaoli Huang, Chen Fang, 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*, pages 12955–12964, 2023. 1