Appendix of Mobile ALOHA: Learning Bimanual Mobile Manipulation using Low-Cost Whole-Body Teleoperation

Author Names Omitted for Anonymous Review. Paper-ID [5]

A. High Five

High Five: The robot base is initialized next to the kitchen island. The robot keeps moving around the kitchen island until a human is in front of it, then high five with the human. Each demo has 2000 steps or 40 seconds, and typically contains 3-4 high fives.

Fig. 1: ***Task Definition of High Five***.

We include the illustration for the *High Five* task in Figure 1. The robot needs to go around the kitchen island, and whenever a human approach it from the front, stop moving and high five with the human. After the high five, the robot should continue moving only when the human moves out of its path. We collect data wearing different clothes and evaluate the trained policy on unseen persons and unseen attires. While this task does not require a lot of precision, it highlights *Mobile ALOHA*'s potential for studying human-robot interactions.

B. Example Image Observations

Figure 2 showcases example images of *Wipe Wine* captured during data collection. The images, arranged sequentially in time from top to bottom, are sourced from three different camera angles from left to right columns: the top egocentric camera, the left wrist camera, and the right wrist camera. The top camera is stationary with respect to the robot frame. In contrast, the wrist cameras are attached to the arms, providing close-up views of the gripper in action. All cameras are set with a fixed focal length and feature auto-exposure to adapt to varying light conditions. These cameras stream at a resolution of 480 × 640 and a frame rate of 30 frames per second.

C. Experiment Details and Hyperparameters of ACT, Diffusion Policy and VINN

We carefully tune the baselines and include the hyperparameters for the baselines and co-training in Table I, II, III, IV, V.

sample prob. from *Mobile ALOHA* data	0.5
sample prob. from *ALOHA* data	0.5

TABLE I: ***Hyperparameters of co-training.***

learning rate	2e-5
batch size	16
# encoder layers	4
# decoder layers	7
feedforward dimension	3200
hidden dimension	512
# heads	8
chunk size	45
beta	10
dropout	0.1
backbone	pretrained ResNet18[1]

TABLE II: ***Hyperparameters of ACT.***

Fig. 2: ***Example Image Observations of Wipe Wine.*** We show the observations from the top camera, left wrist camera and right wrist camera from left to right columns. These images are arranged sequentially in time from top to bottom.

learning rate	1e-4
batch size	32
chunk size	64
scheduler	DDIM[3]
train and test diffusion steps	50, 10
ema power	0.75
backbone	pretrained ResNet18[1]
noise predictor	UNet[2]
image augmentation	RandomCrop(ratio=0.95) & ColorJitter(brightness=0.3, contrast=0.4, saturation=0.5) & RandomRotation(degrees=[-5.0, 5.0])

TABLE III: ***Hyperparameters of Diffusion Policy***.

learning rate	3e-4
batch size	128
epochs	100
momentum	0.9
weight decay	1.5e-6

TABLE IV: ***Hyperparameters of BYOL***, the feature extractor of VINN.

k (nearest neighbour)	selected with lowest validation loss
chunk size	100
state weight	5
camera feature weight	1:1:1 (for front, left and right wrist)

TABLE V: ***Hyperparameters of VINN + Chunking***.

D. Open-Loop Replaying Errors

Figure 3 shows the spread of end-effector error at the end of replaying a 300 steps (6s) demonstration. The demonstration contains a 180 degree turn with radius of roughly 1m. At the end of the trajectory, the right arm would reach out to a piece of paper on the table and tap it gently. The tapping position are then marked on the paper. The red cross denotes the original tapping position, and the red dots are 20 replays of the same trajectory. We observe significant error when replaying the base velocity profile, which is expected due to the stochasticity of the ground contact and low-level controller. Specifically, all replay points are biased to the left side by roughly 10cm, and spread along a line of roughly 20cm. We found our policy to be capable of correcting such errors without explicit localization such as SLAM.

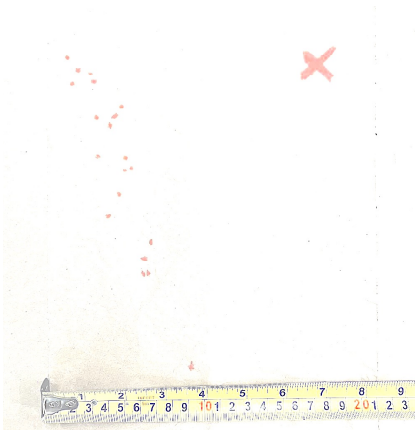

Fig. 3: ***Open-lopp Replay Errors.*** We mark the right arm end-effector position on a piece of paper for the original episode (red cross), and 20 replays of the same episode (red dots).

REFERENCES

[1] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 1, 2

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. URL https://api.semanticscholar.org/CorpusID:3719281. 2

[3] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2