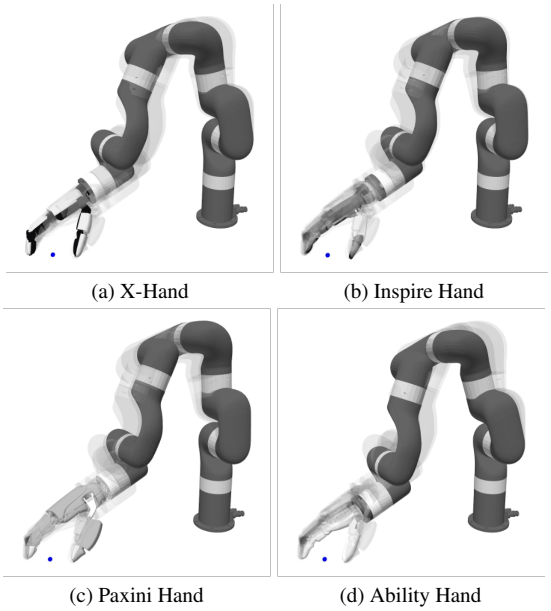
Cross-Hand Latent Representation for Vision-Language-Action Models

7. Appendix

7.1. Latent Visualizations

In addition to Figure 5, we include visualizations of additional hands in Figure 7. This figure illustrates how the same latent representation is decoded across all four hands featured in our main paper. Furthermore, Figure 17 presents a continuous trajectory rendered for all hands, with the X-Hand highlighted for clarity.



(a) X-Hand (b) Inspire Hand

(c) Paxini Hand (d) Ability Hand

Figure 7. **More Latent Visualizations.** Latent decoding results cross embodiment.

7.2. Hardware Setup

Tabletop Scene Description. For the real-world experiments, we use a bimanual arm with tabletop settings. The arms are mounted on the edge of the table. The distance between the arms are 80cm. Each hand is connected to the end effector of the arm with 3D-printed mounts. Figure 8 shows the real-world tabletop scene with a pair of XHand, and figure 9 includes all the dexterous hands we use in our experiments.

Camera Description. We use a single RealSense L515 camera (round-shaped) mounted in front of the bimanual arms as the input view of the policy training. The camera pose is shown in figure 8 and the camera view is in figure 10. The raw resolution of RGB recordings from the camera is 960×540.

Humanoid Scene Description. Similar to xArm, we let G1 stand in front of a table. We use the same camera setting and mount L515 on the chest of G1 to have an egocentric view. Consider the mechanic design of G1,

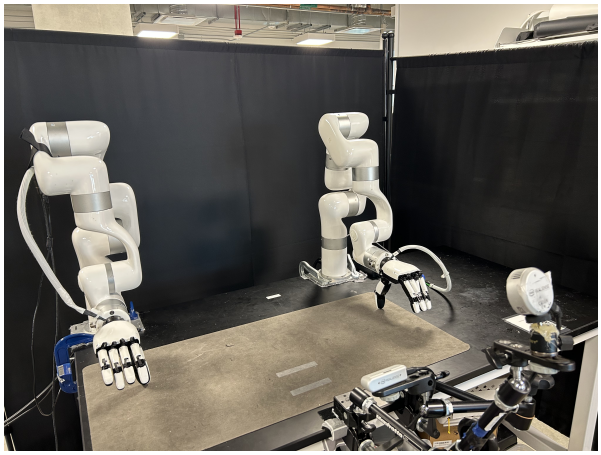


Figure 8. **xArm Camera Setup.** We use a single RealSense L515 camera with the front view. Note that the D435 camera here is not used for XL-VLA.

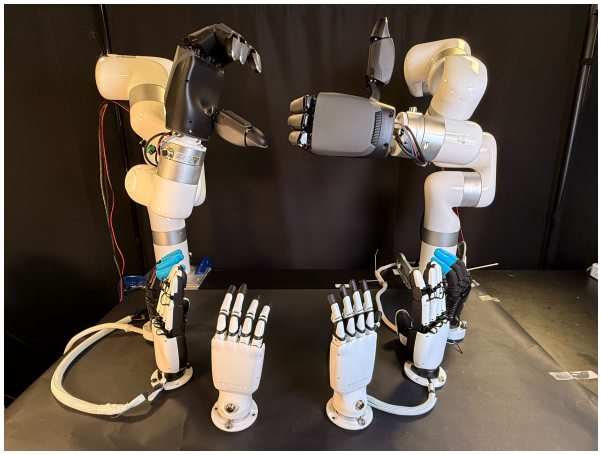


Figure 9. **Dexterous Hands.** We use 4 kinds of hands, with various shapes, scales, degrees of freedom, and actuated joints.

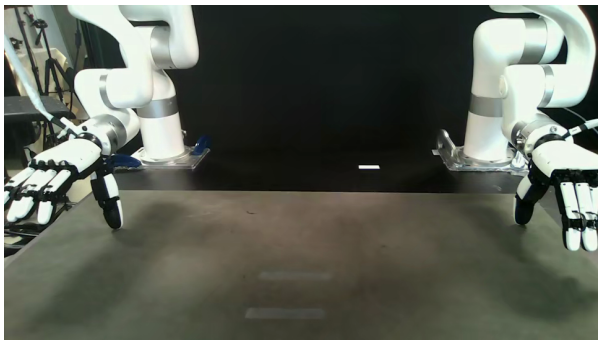


Figure 10. **xArm Camera View.** This is what our camera sees and also the input for XL-VLA and all the baseline methods.

we only use the Inspire hand since it is light. Figure 11 and 12 show the real-world G1 setting.

Figure 11. **G1 Scene.** We mount an L515 camera near the neck of G1 to have an egocentric view.

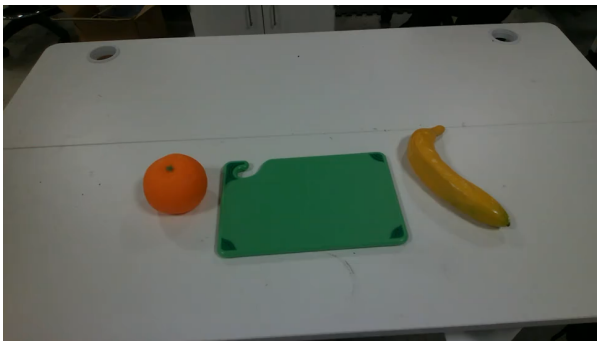


Figure 12. **G1 Egocentric Camera View.**

Object Description. To demonstrate that XL-VLA is capable of doing various manipulation tasks, we use diverse objects in our experiments, most of which are common everyday objects from existing datasets [10, 20]. The objects vary in scale, shape, texture, weight, etc. All the objects used in listed in figure 13.

7.3. Policy Learning Details

xArm Data Collection. We use Apple Vision Pro as the data collection tool [18], shown in figure 14. During data collection, our teleoperator wears the VR headset and get the tracked hand poses and wrist poses from the headset. We use these data to do robot hand retargeting and Inverse Kinematics.

Unitree G1 Data Collection. G1 is standing in front of a table, and we adapt HOMIE [4] and ACE-F [64] to do the upper-body teleoperation. We only use the upper-body system from HOMIE and replace their glove with the MANUS Mocap glove.

Figure 13. **Objects.** We use various everyday objects from existing datasets. They vary in scale, shape, texture, weight, etc., thus requiring the manipulation policy to be robust.

Figure 14. **Apple Vision Por for Data Collection.** Apple Vision Pro is used to track the human teleoperator's hands and wrists. Then the tracked data is processed with retargeting and Inverse Kinematics to control the real-world robots.

Task Visualizations. Using XHand as an example, we show the real-world task visualizations in figure 16. Four of the tasks are also tested on G1.

Model Training. Besides the model training details provided in the main paper, here are some more training details. RGB images is cropped and then resized from 960 $\times$ 540 to 320 $\times$ 240 during data post-processing. When loaded to train XL-VLA, they are resized to 224 $\times$ 224. We use natural language description as the task specification, which is part of the policy condition. The training usually takes around 10 hours for one multi-task policy.

Policy Evaluation. For the real-world evaluation, we do 10 trials for each experiment setting. Among these trials, object positions are randomly initialized while the

Figure 15. **G1 Teleoperation System.** We build the G1 upper-body teleoperation system from HOMIE [4]. We use a pair of MANUS Mocap glove to track the human hand pose.

initial joint position for the robot arm and hand remains the same for the same hand.

For unseen tasks only, we record the partial success rate (PSR). If any of the bimanual robot arm finishes its task and the whole task is failed, the overall success rate is 0.5. For other experiments, we do not use PSR. Only rollout that completes a specified task is count as a success.

7.4. G1 Experiment Results

Table 6 is the numeric results for figure 5.

Method	PF	HB	PS	PoS	Mean
π_0 [6]	0.4	0.6	0.5	0.6	0.525
XL-VLA	**0.7**	**0.9**	**0.9**	**0.8**	**0.825** +57%

Table 6. **G1 Policy Performances.**

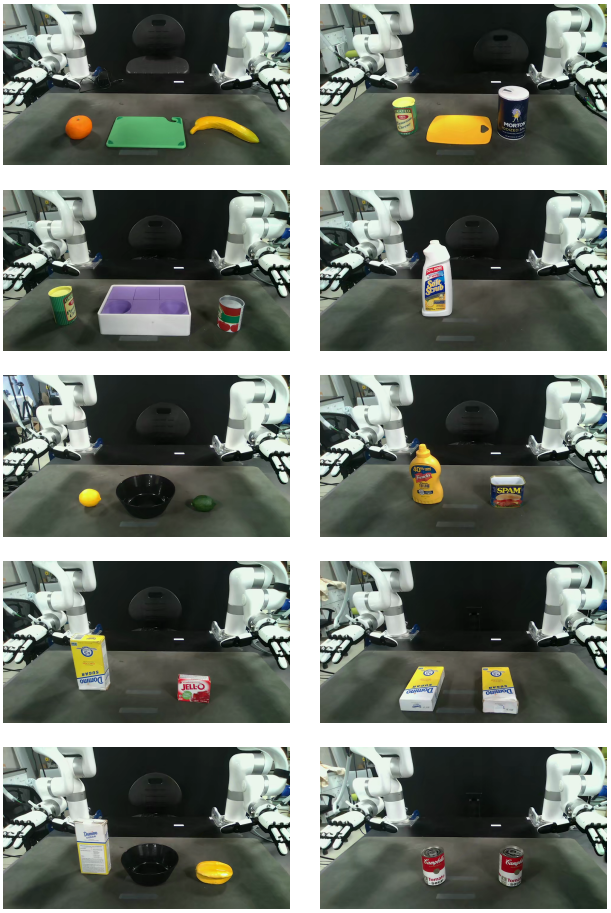


Figure 16. **Task Visualizations.** We design 10 various tasks to test all the models. The tasks require varied manipulation skills and have different difficulties.

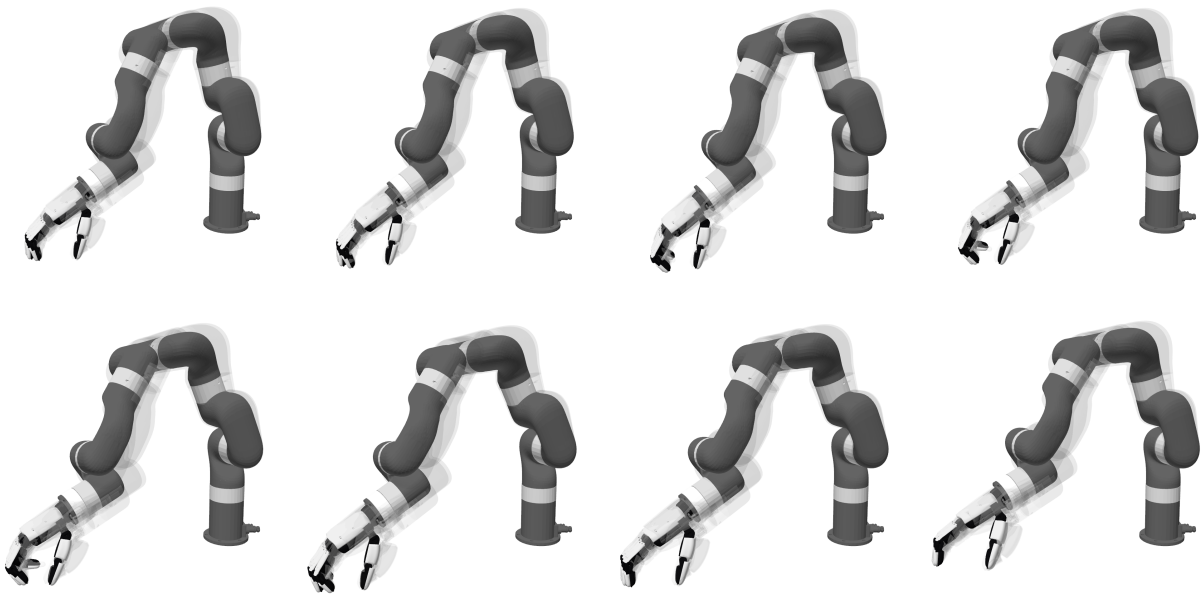


Figure 17. **Latent Visualization of a Grasping Trajectory.** A trajectory is shown here with all the robot hands.