

# Supplementary Material: Open-TeleVision: Teleoperation with Immersive Active Visual Feedback

Anonymous Author(s)

Affiliation

Address

email

## 1 Comparisons with Prior Teleoperation Systems

In this section, we compare our system with various prior teleoperation systems, as listed in Table. 1. We conduct our analysis from two critical perspectives of teleoperation: actuation and perception. The specifics of these comparisons are discussed below.

**Actuation.** Various approaches have been studied for teleoperating robots through human commands, including visual tracking, motion-capture devices, and joint copying through customized hardware. While using motion-capture gloves for teleoperation seems the most intuitive, the commercially available gloves are not only costly but also unable to provide wrist pose estimations. The joint copying method has drawn significant attention recently, following the success of ALOHA[1]. This method offers precise and dexterous control. Historically, this method was considered costly, requiring using an additional pair of identical robotic arms for teleoperation; nonetheless, this issue has been mitigated by the adoption of low-cost exoskeleton devices to transmit commands[2]. Despite their simplicity, joint copying systems are currently limited to using grippers and have not yet been extended to operate multi-finger hands. Conversely, visual tracking employs off-the-shelf hand pose extractors to track finger movements, but relying solely on RGB or RGBD images can lead to noisy and imprecise data. The recent surge in VR technology has led to the development of teleoperation systems that utilize VR tracking. VR headset manufacturers often integrate built-in hand-tracking algorithms that fuse data from diverse types of sensors, including multiple cameras, depth sensors, and IMUs. Hand-tracking data collected through VR devices are generally considered more stable and accurate than self-developed vision-tracking systems, while the latter only utilize a subset of the mentioned sensors (RGB+RGBD[3], Depth+IMU[4], etc.).

**Perception.** While being the other critical component of teleoperation, perception has been considerably less explored than actuation within this field. Most existing teleoperation systems require the operators to directly observe the robot’s hands using their own eyes. While direct viewing provides the operators with depth sensing, leveraging humans’ inherent capability for stereoscopic vision, it restricts the system to be non-remote, necessitating the physical presence of the operator. Some teleoperation systems circumvent this by streaming RGB images, enabling remote control[3, 5]. However, if the operator opts for remote controlling by watching an RGB stream, the benefits of depth sensing provided by the human eye are lost. Despite being capable to provide both remote controlling and depth sensing, these two features are mutually exclusive in these systems. OPEN TEACH[6] merges the two in a mixed-reality fashion, yet it still requires the operator to be in proximity to the robot, otherwise the depth sensing is unavailable. Prior to Open-TeleVision, no system offered both remote control and depth sensing simultaneously: the operator is forced to choose between either direct viewing, which demands physical presence, or RGB streaming, which abandons depth information. Our system is the first to provide both functionalities within a single setup.

Teleop System	Actuation	Hand	Bimanual	Perception	Remote	Depth
OPEN TEACH[6]	VR Tracking	✓	✓	Direct View+RGB	✗	✓
HATO[7]	VR Tracking	✓	✓	Direct View	✗	✓
AnyTeleop[3]	RGB(D) Tracking	✓	✗	Direct View/RGB	✓	✓
Telekinesis[5]	RGB Tracking	✓	✗	Direct View/RGB	✓	✓
Transteleop[4]	IMU+Depth	✓	✗	Direct View	✗	✓
ALOHA[1]	Joint Copy	✗	✓	Direct View	✗	✓
AirExo[2]	Joint Copy	✗	✓	Direct View	✗	✓
GELLO[8]	Joint Copy	✗	✓	Direct View	✗	✓
Mobile ALOHA[9]	Joint Copy	✗	✓	Direct View	✗	✓
DexCap[10]	SLAM+Mocap	✓	✓	Direct View	✗	✓
Open-TeleVision	VR Tracking	✓	✓	Stereo	✓	✓

Table 1: Comparing Open-TeleVision’s capabilities with prior teleoperation systems.

## 2 Discussion of Visual Occlusion

To support our proposed assumption that the unsatisfactory performance observed in *GR-1 Can Sorting* task stems from visual occlusion caused by GR-1’s gripper end-effector, we performed a controlled experiment. In the new experiment, we added color labels to the cans to mitigate the occlusion factor, as depicted in Fig. 1 left. The other settings are identical to those described in the *GR-1 Can Sorting* task in the article. Results are recorded in Table. 2.

Baselines	GR-1 Can Sorting			
	Pick(new)	Pick(old)	Place(new)	Place(old)
Ours	0.97	0.87	1.00	0.60
ResNet	0.90	0.83	0.97	0.50
Mono	0.47	0.73	0.93	0.63

Table 2: Success rate for *GR-1 Can Sorting*, under identical settings and number of trials as outlined in Tab. 1. Columns marked as (old) contain the original results using unlabeled cans, while the columns marked as (new) contain the results of the new experiment using labeled cans.

The results indicate a substantial improvement in the success rate of the placing task across all three baselines, achieved by using labeled cans. *Ours* reached a 100% accuracy rate in the placement, compared to the previous 0.60; notable gains have also been observed in the other baselines, with *ResNet* improving from 0.50 to 0.97, and *textitMono* improving from 0.63 to 0.93. On the other hand, while success rates of picking have also increased for *Ours* and *ResNet*, *Mono* did not exhibit similar improvements. This disparity further validates our claim that a successful can-picking requires spatial information from stereo images.

As with H1 in Sec. 3.2, we performed an experiment to evaluate the model’s generalization capability with *Can Sorting* on GR-1 with labeled cans. Its results are similarly collected from a 4x4 grid (the same as Fig. 5 left) with each cell measuring 3 cm. Generalization results are shown in the heatmap in Fig. 1 right. The results suggest that our model can easily adapt to most of the random locations covered in our experiment, reaching 100% grasping accuracy in nearly all locations on the grid. The results as shown here for *GR-1 Can Sorting* are also notably better than the results as shown in Fig. 5 for *H1 Can Sorting*. The difference may also be

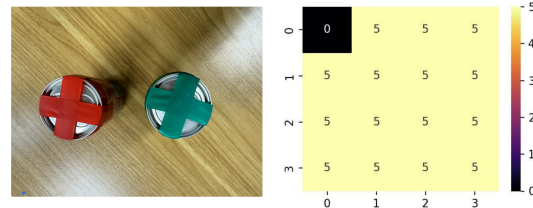


Figure 1: Left: Figure depicting labeled cans. Right: number of successful pickings heatmap with 5 trials at each location.

62 attributed to differences in end-effector morphologies. Grasping a soda can, which requires less  
 63 dexterity and more tolerance, is better suited to grippers than to robotic hands.

### 64 3 Dexterous Hand

65 For H1 robot’s setup, the anthropomorphic hands we used are provided by  
 66 Inspire Robots [11]. A close-up of one of the Inspire Hands is shown in  
 67 Fig. 2. Each hand has five fingers and 12 DoFs, among which 6 are actuated  
 68 DoFs: two actuated DoFs are on the thumb and one on each of the remaining  
 69 fingers. Each non-thumb finger possesses a single actuated revolute joint at  
 70 the metacarpophalangeal (MCP) joint, serving as the entire finger’s actuating  
 71 DoF. The proximal interphalangeal (PIP) joints of these four fingers are driven  
 72 by the MCP joints through linkage mechanisms, adding four underactuated  
 73 DoFs. The thumb is equipped with two actuated DoFs at the carpometacarpal  
 74 (CMC) joint. The thumb’s MCP and interphalangeal (IP) joints are also driven  
 75 by linkage mechanisms, contributing to additional two underactuated DoFs.



Figure 2: Close-up of the Inspire Hand [11].

## 76 4 Experimental Details and Hyperparameters

### 77 4.1 Experimental Details

Tasks	Average Episode Length (s)	Number of Episodes
H1 Can Sorting	93±5	10
GR-1 Can Sorting	61±5	10
Can Insertion	84±7	20
Folding	44±5	20
Unloading	93±6	20

Table 3: Details about collected demonstration data for each task.

78 More experimental details are listed in Tab. 3. All tasks, with the exception of *Can Sorting* (both *H1*  
 79 *Can Sorting* and *GR-1 Can Sorting*), use 20 human demonstrations for training. In contrast, only  
 80 10 demonstrations are used for *Can Sorting*. This choice is primarily due to its repetitive nature:  
 81 each episode consists of 10 (6 for *GR-1 Can Sorting*) individual can-sortings. Consequently, 10  
 82 demonstrations encompass 100 individual sorting rollouts, providing ample data for training.

### 83 4.2 Hyperparameters

84 The hyperparameters employed for training the ACT [1] models are detailed in Table. 4. While the  
 85 majority of these hyperparameters are consistent across all baselines and all tasks, there are a few  
 86 exceptions, including chunk size and temporal weighting. The detailed explanations are as follows.

KL weight	10
chunk size	60
hidden dimension	512
batch size	45
feedforward dimension	3200
epochs	25000
learning rate	5e-5
temporal weighting	0.01

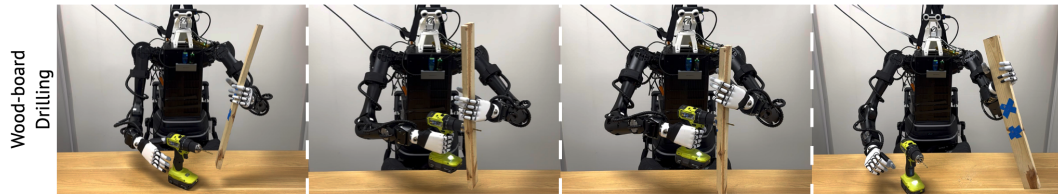
Table 4: Hyperparameters of ACT.

87 The definition of chunk size in the action chunking operation is outlined in the original ACT paper[1].  
88 We use a chunk size of 60 for all tasks, with the exception of *Can Insertion*, in which we use a chunk  
89 size of 100. Using a chunk size of 60 in our setup effectively provides the robot with approximately  
90 one second of memory, correlating with our inference and action frequency of 60Hz. Nonetheless,  
91 we noticed that in *Can Insertion* task, using a larger chunk size, which corresponds to incorporating  
92 more historical actions, proves to be advantageous for the model to perform correct action sequences.

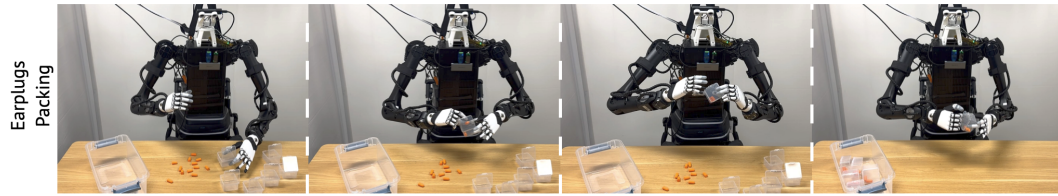
93 The definition of temporal weighting in the temporal aggregation operation is outlined in the original  
94 ACT paper[1], where an exponential weighting scheme  $w_i = \exp(-m * i)$  is employed to assign  
95 weights to actions at different timesteps.  $w_0$  is the weight for the oldest action, adhering to ACT’s  
96 setting.  $m$  is the temporal weighting hyperparameter mentioned in Table. 4. As  $m$  decreases, greater  
97 emphasis is placed on more recent actions, rendering the model more reactive but less steady. We  
98 found that using a temporal weight  $m$  of 0.01 reaches a satisfactory balance between responsiveness  
99 and stability for most tasks. However, for *Unloading* and *Can Sorting* tasks, we adjust this parameter  
100 to cater to their specific needs. For unloading,  $m$  is set as 0.05, ensuring greater stability during  
101 in-hand passing; for *Can Sorting*,  $m$  is set as 0.005, providing quicker movements.

## 102 5 Additional Teleoperation Experiments

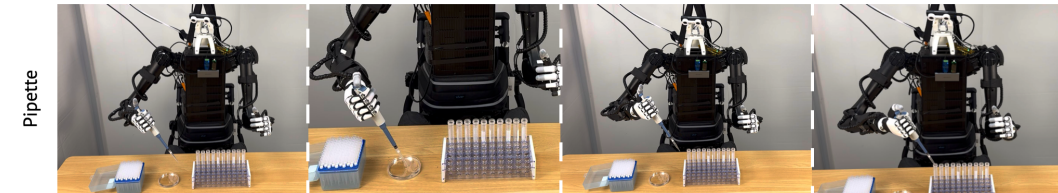
103 In Fig. 3, we include more teleoperation tasks that our system is capable of. The *Wood-board Drilling*  
104 task shows that our system can operate heavy-weight ( $1kg$ ) tools that are designed for humans, thanks  
105 to its compatibility with dexterous hands, and can apply sufficient force to the wood board to drill it  
106 through. Such a task is virtually impossible for the grippers. The *Earplugs Packing* task demonstrates  
107 that our system is dexterous and responsive enough to perform agile bimanual arm-hand coordination.  
108 The *Pipette* task demonstrates that our system is also capable of precise actions. This is also a task  
109 that is extremely hard or impossible for the grippers to achieve, as the usage of a pipette is specialized  
110 for anthropomorphic hands. Even though the motors on H1 humanoid robot are quasi-direct-drive  
111 motors with planetary reducers, which are known to have gear clearance and far less accuracy and  
112 stiffness, our system can still achieve high-precision with human operators in the loop.



(a) The robot holds a wood board of thickness  $2\text{cm}$  with the left hand and uses an electric drill to drill 2 holes on the board. This task requires precise control of the drill trigger using the index finger. Furthermore, our system enables fine control of the hand so that after drilling the first hole, the robot can let the board slide in hand to leave space for the second drilling.



(b) The robot picks randomly placed earplugs on the table and places them into randomly placed latch boxes. The robot needs dexterous bimanual in-hand manipulation and adjustments to properly close the latch box.



(c) The robot utilizes its thumb DoF to control a pipette to transfer liquid from a petri dish to a centrifuge tube. The diameter of the tube is only  $1.5\text{cm}$  so it requires high precision to complete the task.

Figure 3: Additional teleoperation experiments to show our system's reliability and precision for a wide variety of tasks.

## References

- [1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.
- [2] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu. Low-cost exoskeletons for learning whole-arm manipulation in the wild. *arXiv preprint arXiv:2309.14975*, 2023.
- [3] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In *Robotics: Science and Systems*, 2023.
- [4] S. Li, J. Jiang, P. Ruppel, H. Liang, X. Ma, N. Hendrich, F. Sun, and J. Zhang. A mobile robot hand-arm teleoperation system by vision and imu. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10900–10906. IEEE, 2020.
- [5] A. Sivakumar, K. Shaw, and D. Pathak. Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube. *arXiv preprint arXiv:2202.10448*, 2022.
- [6] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto. Open teach: A versatile teleoperation system for robotic manipulation, 2024.
- [7] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik. Learning visuotactile skills with two multifingered hands. *arXiv:2404.16823*, 2024.
- [8] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. *arXiv preprint arXiv:2309.13037*, 2023.
- [9] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.
- [10] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.
- [11] Inspire Robots, Dexterous Hands, 2024, [www.inspire-robots.store/collections/the-dexterous-hands](http://www.inspire-robots.store/collections/the-dexterous-hands), [Online; accessed Jun. 2024].