

# Bunny-VisionPro: Real-Time Bimanual Dexterous Teleoperation for Imitation Learning

Runyu Ding<sup>1,2\*</sup>, Yuzhe Qin<sup>1\*</sup>, Jiyue Zhu<sup>1\*</sup>, Chengzhe Jia<sup>1</sup>,  
Shiqi Yang<sup>1</sup>, Ruihan Yang<sup>1</sup>, Xiaojuan Qi<sup>2</sup> and Xiaolong Wang<sup>1</sup>

**Abstract**—Teleoperation is a crucial tool for collecting human demonstrations, but controlling robots with bimanual dexterous hands remains a challenge. Existing teleoperation systems struggle to handle the complexity of coordinating two hands for intricate manipulations. We introduce Bunny-VisionPro, a real-time bimanual dexterous teleoperation system that leverages a VR headset. Unlike previous vision-based teleoperation systems, we design novel low-cost devices to provide haptic feedback to the operator, enhancing immersion. Our system prioritizes safety by incorporating collision and singularity avoidance while maintaining real-time performance through innovative designs. Bunny-VisionPro outperforms prior systems on a standard task suite, achieving higher success rates and reduced task completion times. Moreover, the high-quality teleoperation demonstrations improve downstream imitation learning performance, leading to better generalizability. Notably, Bunny-VisionPro enables imitation learning with challenging multi-stage, long-horizon dexterous manipulation tasks, which have rarely been addressed in previous work. Our system’s ability to handle bimanual manipulations while prioritizing safety and real-time performance makes it a powerful tool for advancing dexterous manipulation and imitation learning. Our web page is available at [https://dingry.github.io/projects/bunny\\_visionpro](https://dingry.github.io/projects/bunny_visionpro).

## I. INTRODUCTION

Playing a Virtual Reality (VR) game is an immersive and intuitive experience where your hand and arm movements seamlessly translate into the action of a virtual character. Now, imagine controlling a bimanual robot in the real world with the same ease: operators use their own movements to guide the robot’s motion, just as they would in a VR game. This paradigm shift in robot teleoperation opens up exciting possibilities for more intuitive and accessible human-robot interaction. Recent advancements in VR technology, such as the Apple Vision Pro, have made this concept a possibility.

Yet, translating this concept into a practical teleoperation system presents significant challenges due to the intricate motion required to perform human-like manipulations. The complexity is further amplified for high Degree-of-Freedom (DoF) hand-arm systems, where the operator must coordinate two arms and hands to execute tasks requiring spatiotemporal synchronization. Achieving responsive control is crucial, as delays can lead to imprecise robot motion [1]. Besides, ensuring safety by mitigating risks such as collision and singularity adds another layer of complexity [2].

In this paper, we introduce Bunny-VisionPro, a novel real-time bimanual dexterous teleoperation system. As illustrated

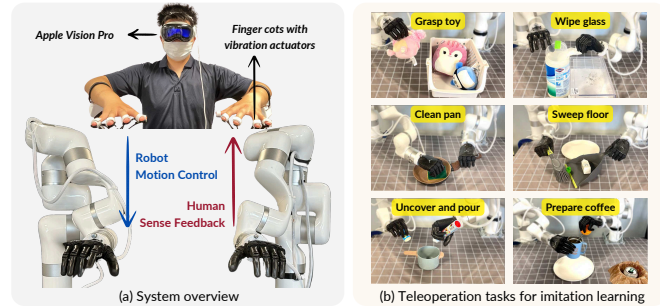


Fig. 1: **System Overview and Task Suits.** (a) Hand poses captured by Apple Vision Pro are converted into robot motion control commands for real-time teleoperation. The robot provides sensory feedback, including vision and touch, to operators via Vision Pro and actuator-equipped finger cots. (b) We design diverse short-horizon (left-column) and long-horizon tasks (right-column) to evaluate teleoperation performance and its application for imitation learning.

in Fig. 1, our system utilizes Vision Pro’s tracking capabilities to capture operator hand movements and translate these movements into precise robotic commands. It addresses both arm and dexterous hand control, prioritizing safety and real-time performance. Crucially, Bunny-VisionPro features three innovative modules: (i) The arm motion control module, which can handle robot singularity and collision avoidance in real time without the need for high-end GPUs. This module ensures the safe and smooth operation of robot arms, even in complex bimanual manipulation scenarios. (ii) The dexterous hand retargeting module, which accurately maps human finger movements to the robotic hand. This module features its unique ability to perform retargeting for robots with loop joints, such as four-bar linkages, in real time. (iii) The haptic feedback module, which uses low-cost Eccentric Rotating Mass (ERM) actuators (\$1.2 each) to provide tactile sensations. This novel design, paired with a VR headset, offers a more immersive and realistic experience as the robot interacts with its environment, creating a sensation that the operator is the robot itself.

The advanced capabilities of Bunny-VisionPro enable the collection of high-quality demonstrations for dexterous, bimanual, and long-horizon imitation learning tasks. By leveraging the intuitive and responsive teleoperation system, operators can perform intricate manipulations and collect diverse data for learning algorithms, enabling them to learn and generalize to new scenarios effectively.

We evaluate our Bunny-VisionPro system on the Telekine-

Runyu Ding was an intern at UC San Diego during the project  
Jiyue Zhu is the corresponding author. [jiz226@ucsd.edu](mailto:jiz226@ucsd.edu)

\* indicates equal contribution

<sup>1</sup>UC San Diego, <sup>2</sup>The University of Hong Kong

sis [3] benchmark, achieving 11% higher success rates and reducing task completion time by 45% compared to prior systems. These enhancements are especially pronounced in multi-stage, long-horizon tasks. Imitation learning policies trained on demonstrations collected by our system show a 20% improvement in generalization on novel poses and unseen objects compared to using data from previous work [4]. These results demonstrate the system's superior performance in executing complex bimanual manipulation tasks and its effectiveness in collecting high-quality data.

## II. RELATED WORK

**Teleoperation with Gripper.** Classical teleoperation methods can be categorized into two main approaches based on their control objectives. The first approach, exemplified by ALOHA [5], [6], [7], uses joint-space mapping [8] within leader-follower setups. Although enabling impressive bimanual manipulation, this approach is robot-specific, requiring kinematic equivalence between the leader and follower robots [9]. It also places the burden of managing collision and singularity on the human operator. The second approach prioritizes end-effector control [10], with arm joint positions calculated using inverse kinematics. Various input devices, such as motion capture systems [11], [12], inertia sensors [13], and VR controllers [14], [15], [16], [17] have been utilized. Nevertheless, these systems often employ simple 1 or 2 DoF grippers, which limits their dexterity. In contrast, our work aims to teleoperate high-DoF hand-arm systems for complex manipulation tasks.

**Dexterous Teleoperation.** Teleoperating dexterous hands is challenging due to their high DoF and complex kinematics. Glove-based systems [18], [19], [20], [21], like the MANUS glove [22] used by Tesla Bot, can track the operator's finger movements but are costly and require specific hand sizes. Recent vision-based approaches, such as AnyTeleop [4], facilitate dexterous hand-arm teleoperation using cameras [3], [23], [24], [25] or VR headsets [26], [27], [28]. However, AnyTeleop requires sophisticated GPU processing for arm motion computation and is primarily designed for a single arm. A concurrent work [29] controls the Ability hand with reduced DoF using grip buttons, sacrificing dexterous finger gaiting to avoid retargeting latency. In contrast, our retargeting modules can handle the Ability hand's four-bar linkage structure at 300Hz with one CPU core, enabling full dexterity. Our system also integrates collision and singularity avoidance while maintaining real-time performance.

**Imitation Learning from Demonstration.** Imitation learning enables robots to mimic human behaviors through expert guidance. Pioneering research utilizing deep learning [30], [31], [32], [33] develop policies to generate robot control commands based on image [5], [34], [35], [36], [37], [38] and point cloud data [39], [40], [41], [42], and further marking progress in bimanual systems [18], [5], [6], [43]. Additionally, recent works incorporate tactile data [44], [29] to enrich the sensory data pool for robot learning. Given that demonstration collection is labor-intensive yet highly critical for effective imitation learning, our system offers a

compelling solution that streamlines the collection of high-quality bimanual dexterous data, greatly enhancing accuracy and generalization across various learning algorithms.

## III. TELEOPERATION SYSTEM

### A. Overview

Bunny-VisionPro is a modular teleoperation system using a VR headset's hand and wrist tracking capabilities to control a high-DoF bimanual robot, as shown in Fig. 2. The system consists of three decoupled components: hand motion retargeting, arm motion control, and human haptic feedback.

The hand motion retargeting module maps the operator's finger poses to the robot's dexterous hands, enabling intuitive dexterous manipulation. Concurrently, the arm motion control module uses the operator's wrist poses as input to compute joint angles for the robot arms while considering collision avoidance and singularity handling. The human haptic feedback module converts the tactile readings from sensors mounted on the robot's hands into actuation signals for the wearable Eccentric Rotating Mass (ERM) actuators on the operator's hands. This provides the operator with real-time haptic feedback, enhancing their sense of presence and allowing for more precise manipulation.

Coordinating bimanual robots during teleoperation also requires maintaining the distance between two robot hands to match the distance between two operators' hands, ensuring natural, coordinated movements. However, the initial distance between the robot hands may not match that of the human hands. To address this, we design several initialization modes that align the robot's hands with the operator's hands on the fly when the robot starts to move. Different tasks may benefit from different initialization modes. This step creates a consistent starting point, enabling intuitive and efficient bimanual task execution throughout the teleoperation process.

For communication, we stream hand pose results from Vision Pro to the computer via [45]. The modular architecture allows for better extensibility and enables each module to run in a separate computation process, preventing latency accumulation in the system and ensuring real-time control.

### B. Robot Hand Motion Retargeting

The hand motion retargeting module translates human finger poses into corresponding robot joint positions. This is achieved by minimizing the difference between the fingertip keypoint vectors [46] of the human hand and the robotic hand, formulated as an online optimization problem:

$$L_{\text{hand}} = \sum_{i=1}^N \|\alpha v_i - \text{FK}_i(q)\|_2^2 + \beta \|\Delta q\|_2^2 \quad \text{s.t.} \quad q^l \leq q \leq q^u. \quad (1)$$

The objective function  $L_{\text{hand}}$  consists of two terms: (i) The first term minimizes the distance between the scaled human hand keypoint vectors  $\alpha v_i$  and the robot hand keypoint vectors computed by forward kinematics function  $\text{FK}_i(q)$ , where  $v_i$  represents the  $i^{\text{th}}$  fingertip keypoint vector,  $N$  is the total number of vectors,  $\|\cdot\|_2$  denotes Euclidean norm,  $\alpha$  is the scaling factor to adjust for size differences between the human and robot hand and  $q$  is the joint positions of the robot hand bounded by the lower and upper limit,  $q^l$  and  $q^u$ . (ii)

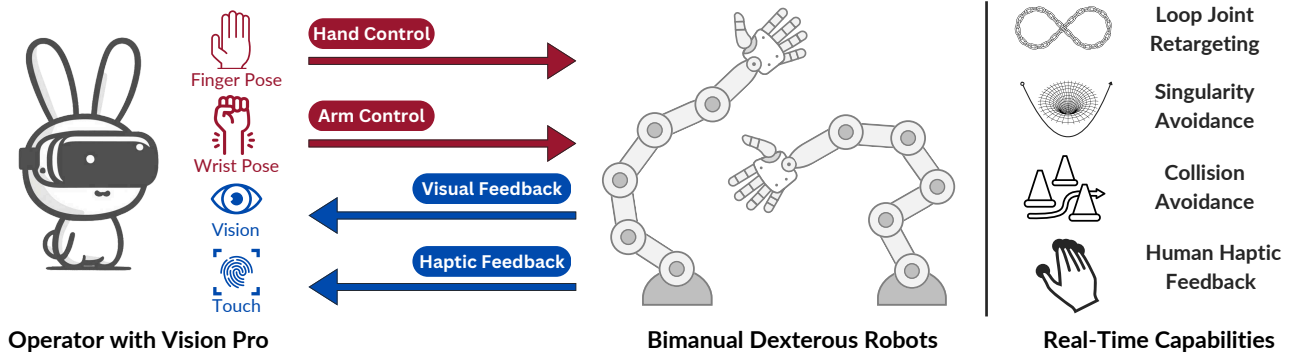


Fig. 2: **Teleoperation System.** The operator controls the robot hand and arm using finger and wrist poses, respectively. The system’s visual and haptic feedback, combined with its four real-time capabilities, provides an intuitive and immersive VR experience for the operator.

The second term enforces temporal smoothness by penalizing large joint position changes  $\Delta q$  between consecutive frames, with weight  $\beta$ . The optimization variable is  $q$  and the objective is solved using Sequential Quadratic Programming (SQP) [47], [48], a gradient-based method that iteratively solves quadratic approximations of the problem.

Our work uniquely addresses the real-time handling of loop joints, common in dexterous hands, where the number of joints exceeds the actual DoF. For a robot with  $n$  joints, the positions of  $n - k$  passive joints depend on the other  $k$  active joint, represented as:

$$q_j = c_j(q_1, q_2, \dots, q_k), \text{ for } j \in \{k + 1, \dots, n\}, \quad (2)$$

where the passive joint  $j$ ’s position  $q_j$  is not independent but constrained by function  $c_j$ .  $n$  is the total number of joint positions. Instead of adding equality constraints to the optimization problem in Eq. (1), which can lead to instability and increased computation time due to the highly non-linear forward kinematic function  $FK(q)$ , we reformulate the problem into a reduced dimension where the optimization variable has  $k$  dimensions (active joints) instead of  $n$  (total joints). Specifically, in the forward pass of optimization, the positions of passive joints are computed as Eq. (2). In the backward pass, the gradients of the passive joints are backpropagated to the active joints and added to their original gradients. Thereby, the gradient for the active joint  $i$  is:

$$\text{grad}_i = \frac{\partial L_{\text{hand}}}{\partial q_i} + \sum_{j=k+1}^n \frac{\partial c_j}{\partial q_i} \frac{\partial L_{\text{hand}}}{\partial q_j}. \quad (3)$$

This approach eliminates the need for SQP to handle affine approximations of the constraints, making the problem more tractable. Experimental results in Tab. I show a **10.2x speedup** when using this method to solve the four-bar linkage structure (a type of loop joint) in the Ability Hand.

### C. Robot Arm Motion Control

Arm motion control involves computing the joint trajectories of the robot arm based on the wrist pose detected by Vision Pro. Traditional approaches often rely on closed-form solutions for Inverse Kinematics (IK) and handle constraints such as singularity by exploring the null space [49]. However, for typical 7-DoF arms, the 1-DoF null space may not

provide enough flexibility. Modern approaches formulate the problem as an optimization task [50], similar to retargeting. It relaxes the rigid constraints of IK to accommodate additional factors. Inspired by these works, we have developed a unified and efficient optimization objective  $L_{\text{arm}}$  in Eq.(4) that integrates IK, collision avoidance, and singularity management for real-time motion control.

$$L_{\text{arm}} = L_{\text{ik}} + L_{\text{sin}} + L_{\text{col}}, \quad (4)$$

The first term concerns the IK objective:

$$L_{\text{ik}} = \beta_{\text{pos}} \|p_{\text{ee}} - p_{\text{wrist}}\|_2 + \beta_{\text{rot}} \cdot \arccos(2\langle q_{\text{ee}}, q_{\text{wrist}} \rangle^2 - 1). \quad (5)$$

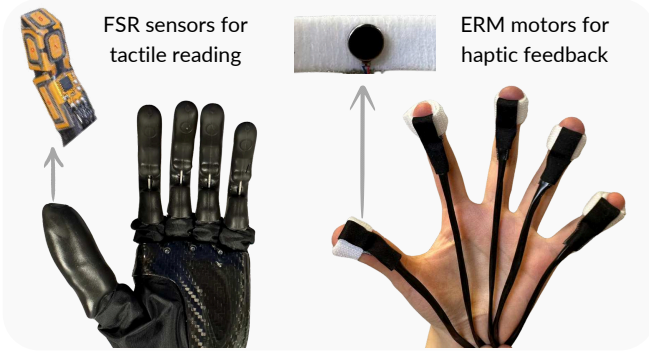
Here,  $p_{\text{ee}}$  and  $p_{\text{wrist}}$  represent the robot end effector and human wrist position, respectively, while  $q_{\text{ee}}$  and  $q_{\text{wrist}}$  are their corresponding quaternions.  $\langle \cdot, \cdot \rangle$  means inner product. The weights  $\beta_{\text{pos}}$  and  $\beta_{\text{rot}}$  allow for balancing the errors between position and rotation. Additionally, the objectives for singularity and collision avoidance are as follows:

$$L_{\text{sin}} = \begin{cases} 1 - \frac{\sqrt{\det[\mathbf{J}\mathbf{J}^T]}}{\lambda}, & \text{if } s_0 < s_{\text{low}} \\ 0, & \text{otherwise} \end{cases}, \quad (6)$$

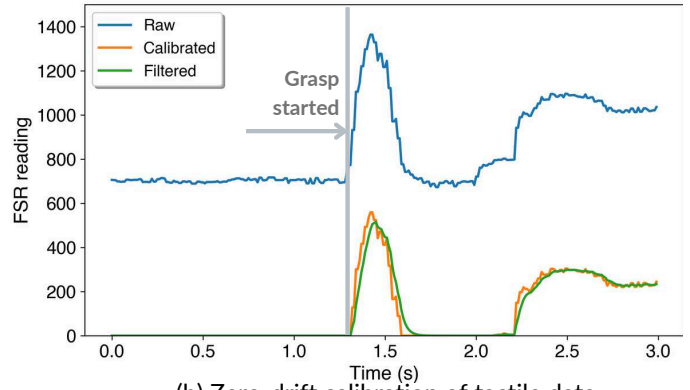
$$L_{\text{col}} = \frac{1}{\sum_{i=1}^m \sum_{j=1}^m \text{dist}(e_i, e_j) \mathbb{1}(e_i, e_j) + \epsilon}. \quad (7)$$

In the singularity avoidance objective  $L_{\text{sin}}$ ,  $\mathbf{J}$  represents the spatial Jacobian matrix,  $s_0$  is the smallest singular value,  $s_{\text{low}}$  is a conditional trigger for penalizing singularity, and  $\lambda$  acts as a temperature factor. Theoretically, both  $s_0$  and the manipulability index  $\sqrt{\det[\mathbf{J}\mathbf{J}^T]}$  can measure singularity [51], and the robot is nearing singularity when either value is low. We opt for the manipulability index as the objective due to its suitability for gradient-based optimization while using  $s_0$  as the condition. Thus,  $L_{\text{sin}}$  only takes effect when the robot approaches singularity.

For collision detection, the GJK algorithm [52] is often employed but can be time-intensive even with convex meshes. To efficiently compute the self-collision cost  $L_{\text{col}}$ , we model each robot link as a collection of  $m$  spheres. The function  $\text{dist}(\cdot, \cdot)$  calculates the Euclidean distance between spheres  $e_i$  and  $e_j$ , while  $\mathbb{1}(\cdot, \cdot)$  serves as an indicator function to evaluate the necessity of collision checking, with  $\epsilon$  provides numerical stability. Collisions are disregarded



(a) Human haptic feedback from robot tactile readings



(b) Zero-drift calibration of tactile data

Fig. 3: **Human Haptics Feedback Device.** Tactile readings from FSR sensors located in the robot’s fingertips undergo calibration and low-pass filtering. The processed tactile signals then drive ERM motors to deliver touch feedback to human.

between spheres within the same link or between spheres of a parent and its direct child. This simplification not only accelerates computation but also renders the distance function differentiable.

#### D. Human Haptic Feedback Device

Effective human manipulation relies on the integration of both visual and tactile feedback. However, many vision-based teleoperation systems [4], [23], [26], [28], [3] neglect haptic feedback. To address this limitation, we have developed a cost-effective haptic feedback system using Eccentric Rotating Mass (ERM) actuators (Fig. 3(a)). This system first processes tactile signals from robot hands (Step I) and then drives vibration motors to simulate tactile sensations (Step II). Despite its low cost, our system enables the operator to perceive and respond to the environment more intuitively with a more immersive experience, leading to improved manipulation performance.

**Step I: Tactile Signal Processing.** The Ability hand (Fig. 3(a)) uses Force-Sensitive Resistors (FSRs) to measure finger pressure. However, FSR sensors suffer from imprecision and zero-drift problems [53], [54], which are aggravated by deformable wrapping materials. To address this, we perform zero-drift calibration by recording baseline FSR readings at various joint positions and subtracting interpolated baseline values from real-time readings during operation. This efficient calibration can be performed autonomously each time teleoperation is initiated. Furthermore, we apply a low-pass filter to reduce noise and smooth the tactile data. As shown in Fig. 3(b), these signal-processing techniques significantly improve the quality and reliability of the tactile feedback.

**Step II: Vibration Motor Driving.** In this step, we convert the processed tactile signals to ERM actuator vibrations using an ELEGOO UNO board. Since ERM motors require constant input voltage, we employ Pulse-Width Modulation (PWM) to simulate continuous haptic strength by controlling ERM vibration intensity through modulated pulse width. To further enhance the stability of the haptic signals and the system’s robustness, we add a Bipolar Junction Transistor (BJT) between each ERM actuator and its corresponding

TABLE I: **Profiling Results.** We measure module latencies in our system.

| Module                                 | Time (ms) |
|----------------------------------------|-----------|
| <i>Retargeting:</i>                    |           |
| w/o loop joints                        | 2.54      |
| Loop joints (baseline, as constraints) | 34.98     |
| Loop joints (ours)                     | 3.43      |
| <i>Motion Control:</i>                 |           |
| IK only                                | 0.74      |
| + Collision                            | 7.85      |
| + Singularity                          | 10.42     |
| + Collision + Singularity              | 15.93     |
| Haptic Feedback (PWM)                  | 0.04      |

TABLE II: **Teleoperation Results in Telekinesis Benchmark.** 10 single-arm tasks are evaluated.

| Task                        | Telekinesis [3] | AnyTeleop [4] | Ours         |
|-----------------------------|-----------------|---------------|--------------|
| Pickup Box Object           | 9/10            | <b>10/10</b>  | <b>20/20</b> |
| Pickup Fabric Toy           | 9/10            | <b>10/10</b>  | <b>20/20</b> |
| Box Rotation                | 6/10            | 6/10          | <b>18/20</b> |
| Scissor Pickup              | 7/10            | <b>8/10</b>   | 15/20        |
| Cup Stack                   | 6/10            | 9/10          | <b>20/20</b> |
| Two Cup Stacking            | 3/10            | 7/10          | <b>19/20</b> |
| Pouring Cubes onto Plate    | 7/10            | 7/10          | <b>20/20</b> |
| Cup Into Plate              | 8/10            | <b>10/10</b>  | <b>20/20</b> |
| Open Drawer                 | 9/10            | <b>10/10</b>  | <b>20/20</b> |
| Open Drawer & Pickup Object | 6/10            | 9/10          | <b>20/20</b> |

PWM pin. This ensures that the haptic feedback of each ERM motor is individually adjusted, only activated during the pulse width, and resistant to circuit noise.

## IV. SYSTEM EVALUATION

We validate our system through: (1) Profiling analysis to verify real-time capabilities (Sec. IV-A); (2) Performance comparison with advanced teleoperation systems on real-robot dexterous tasks (Sec. IV-B); (3) User study with untrained operators to assess haptic usability (Sec. IV-C).

### A. Profiling Results

We profile module latencies in our system using an i7-12700KF CPU (see Tab. I). Our method retargets hand motion on the Ability hand, processing loop joints as efficiently as methods without loop joints and much faster than constraint-based approaches (baseline in Tab. I). Our

TABLE III: **Teleoperation System Evaluation.** We measure collection success rate (Success), episode length (EpLen), collection time in minutes (Time), and joint position changes of the arms ( $\Delta Q_{pos}$ ).

| Task                | Description                                             | Arm    | System     | Success ( $\uparrow$ ) | Time ( $\downarrow$ ) | EpLen ( $\downarrow$ )       | $\Delta Q_{pos}$ ( $\downarrow$ ) |
|---------------------|---------------------------------------------------------|--------|------------|------------------------|-----------------------|------------------------------|-----------------------------------|
| Short-horizon Tasks |                                                         |        |            |                        |                       |                              |                                   |
| Grasp Toy           | Grasp toy and place in box.                             | single | AnyTeleop+ | <b>50 / 54</b>         | 28.9                  | 145 $\pm$ 24                 | 0.77                              |
|                     |                                                         |        | Ours       | 50 / 55                | <b>24.3</b>           | <b>142<math>\pm</math>21</b> | <b>0.35</b>                       |
| Clean Pan           | Grasp panhandle and sponge to clean the pan.            | dual   | AnyTeleop+ | 50 / 78                | 74.6                  | 282 $\pm$ 23                 | 0.98                              |
|                     |                                                         |        | Ours       | <b>50 / 58</b>         | <b>30.4</b>           | <b>186<math>\pm</math>17</b> | <b>0.58</b>                       |
| Uncover & Pour      | Grasp container, remove lid, and pour into bowl.        | dual   | AnyTeleop+ | 50 / 77                | 63.2                  | 189 $\pm$ 21                 | 1.26                              |
|                     |                                                         |        | Ours       | <b>50 / 63</b>         | <b>41.5</b>           | <b>183<math>\pm</math>15</b> | <b>0.86</b>                       |
| Long-horizon Tasks  |                                                         |        |            |                        |                       |                              |                                   |
| Wipe Glass          | Pull out wipes to clean glass                           | dual   | AnyTeleop+ | 30 / 53                | 69.0                  | 517 $\pm$ 51                 | 1.44                              |
|                     |                                                         |        | Ours       | <b>30 / 41</b>         | <b>32.2</b>           | <b>360<math>\pm</math>35</b> | <b>0.68</b>                       |
| Sweep Floor         | Sweep and dump trash using dustpan and broom            | dual   | AnyTeleop+ | 30 / 61                | 129.4                 | 786 $\pm$ 118                | 1.99                              |
|                     |                                                         |        | Ours       | <b>30 / 40</b>         | <b>45.4</b>           | <b>570<math>\pm</math>77</b> | <b>1.30</b>                       |
| Prepare Coffee      | Pour coffee into bottle, stir and hand it over to human | dual   | AnyTeleop+ | 30 / 52                | 95.3                  | 705 $\pm$ 42                 | 1.87                              |
|                     |                                                         |        | Ours       | <b>30 / 44</b>         | <b>55.2</b>           | <b>576<math>\pm</math>33</b> | <b>1.05</b>                       |

motion control module, including singularity and collision avoidance, operates in real-time at  $> 60$  Hz.

### B. Real Robot Teleoperation Experiments

**Task Suits.** The Teleskinesis [3] benchmark comprises ten single-arm-hand manipulation tasks, as listed in Tab. II. To further gauge our system’s capability primarily for bimanual dexterous manipulation, we design three short-horizon tasks – *grasping toy*, *cleaning pan* and *uncovering and pouring* – and three long-horizon tasks – *wiping glass*, *sweeping floor* and *preparing coffee*, as shown in Fig. 1. These tasks, except for *grasping toy*, necessitate fine-grained coordination between two hands.

**Real-world Robot Setup.** Our bimanual dexterous system consists of two UFactory xArm-7 arms, each equipped with a 6-DoF Ability hand, resulting in a 26-DoF system. Each hand incorporates 30 tactile sensors distributed across five fingertips. For data collection, two RealSense L515 cameras are positioned at the front and the top of the robot’s workspace to capture adequate visual observations.

**Baselines.** We assess our system on the Teleskinesis benchmark against Teleskinesis [3] and AnyTeleop [4], using the results reported in the original papers for comparison. For our custom tasks, to ensure a fair comparison, we integrate AnyTeleop with our Vision-Pro-based hand tracking to create AnyTeleop+ and use an identical real-robot setup as our Bunny-VisionPro.

**Teleoperation Results.** As shown in Tab. II, Bunny-VisionPro matches or surpasses baseline methods in 9 out of 10 tasks, evidencing its system precision and robustness for hand manipulation teleoperation. However, in the *scissor pickup* task, our system underperforms slightly due to the limited DoF in the Ability hand, which makes it challenging to insert fingers into the scissor handles.

In self-designed manipulation tasks (Tab. III), our system outperforms AnyTeleop+ [4], achieving 11% higher success rate with 45% task completion time. This performance confirms its precision in replicating human actions and its responsiveness and efficiency. Additionally, Bunny-VisionPro demonstrates enhanced stability and robustness, reducing episode lengths by 19% with lower variability. In

contrast, AnyTeleop+ struggles with complex trajectories and large end-effector pose changes, often leading to aggressive joint movements and unpredictable, potentially unsafe robot control, as evidenced by a 43% increase in arm joint position changes during tasks. Besides, our system’s advantages are particularly pronounced in bimanual long-horizon tasks, where it excels in real-time, fine-grained coordination between two hands, leading to a better teleoperation experience.

### C. User Study of Haptic Feedback

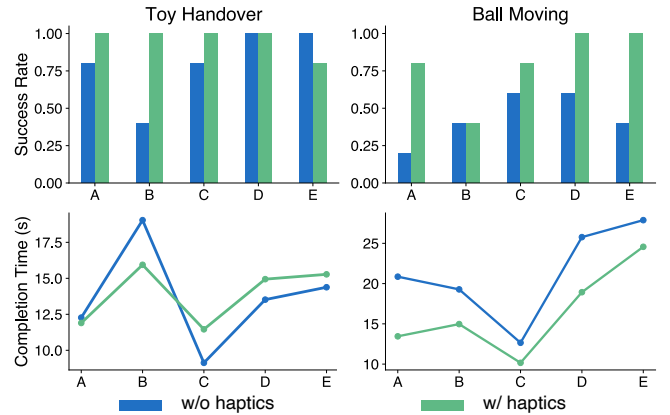


Fig. 4: **User Study.** It evaluates the impact of haptic feedback on success rates and time efficiency.

We conduct a user study on human haptic feedback with five untrained operators. Participants perform *toy handover* and *ball moving* tasks, five trials for each task, with task success rates and average completion time recorded. As shown in Fig. 4, haptic feedback greatly improves success rate and efficiency in the ball-moving task where vision is partially occluded and touch sensing is crucial. This boost stems from the enhanced interaction awareness between the robot and the object, which prevents insufficient or excessive pressure, thus promoting safer and quicker operations. The toy-handover task also sees slight improvements in success rate. This may be attributed to the enhanced confidence in teleoperation reported by operators when using haptic feedback, regardless of visual obstacles. More haptics comparisons are on the project page.

TABLE IV: **Imitation Learning for Real-World Tasks.** The Success Rate (SR) is based on 10 trials, with SR-SG indicating success rates for spatial generalization and SR-U for unseen objects.

| Task           | Data Collection System | ACT [5]     |             |             | Diffusion Policy [34] |             |             | DP3 [40]     |             |             |
|----------------|------------------------|-------------|-------------|-------------|-----------------------|-------------|-------------|--------------|-------------|-------------|
|                |                        | SR          | SR-SG       | SR-U        | SR                    | SR-SG       | SR-U        | SR           | SR-SG       | SR-U        |
| Grasp toy      | AnyTeleop+             | <b>8/10</b> | <b>6/10</b> | <b>7/10</b> | 4/10                  | 3/10        | 3/10        | <b>10/10</b> | 5/10        | 5/10        |
|                | Ours                   | <b>8/10</b> | 4/10        | 5/10        | <b>7/10</b>           | <b>6/10</b> | <b>4/10</b> | <b>10/10</b> | <b>6/10</b> | <b>7/10</b> |
| Clean pan      | AnyTeleop+             | 7/10        | 3/10        | 2/10        | 5/10                  | <b>4/10</b> | 1/10        | 6/10         | 1/10        | 4/10        |
|                | Ours                   | <b>9/10</b> | <b>6/10</b> | <b>5/10</b> | <b>7/10</b>           | <b>4/10</b> | <b>7/10</b> | <b>8/10</b>  | <b>4/10</b> | <b>5/10</b> |
| Uncover & pour | AnyTeleop+             | 2/10        | 0/10        | 1/10        | 3/10                  | <b>2/10</b> | 2/10        | 5/10         | 4/10        | 3/10        |
|                | Ours                   | <b>8/10</b> | <b>4/10</b> | <b>8/10</b> | <b>6/10</b>           | <b>2/10</b> | <b>3/10</b> | <b>7/10</b>  | <b>5/10</b> | <b>7/10</b> |

## V. IMITATION LEARNING

To assess the quality of demonstrations collected by our system compared to AnyTeleop+, we evaluate their impact on imitation learning. We train policies using three methods: ACT [5], Diffusion Policy [34], and DP3 [40], then test their generalization to unseen scenarios. This allows us to measure data quality independently of specific learning algorithms. Besides, we investigate the effectiveness of tactile data in imitation learning.

### A. Main Results.

As shown in Tab. IV, Bunny-VisionPro significantly outperforms AnyTeleop+ by an average of 22% success rate across three tasks using three different policies. This showcases the superior quality of the demonstrations collected by our system. Our system’s effectiveness is particularly evident in bimanual tasks, *i.e.* *cleaning pan* and *uncovering and pouring*. One possible reason for this improvement is that our system handles the alignment of bimanual robots delicately, enabling the collection of demonstrations with more natural robot motions.

**Generalization.** We examine data quality impact on policy generalization by comparing policies trained on AnyTeleop+’s and Bunny-VisionPro’s data. We assess spatial generalization (SR-SG) with random object poses and manipulation of unseen objects (SR-U) with varying shapes, sizes, and colors. Across three policies trained on each dataset, those using Bunny-VisionPro’s data show consistent gains of 14% SR-SG and 26% SR-U (Tab. IV). It suggests our system enables high-quality, trajectory-stable data collection, which benefits imitation generalization.

**Long-horizon Tasks.** For complicated, multi-stage tasks, the policy achieves notable success rates – 73% for sub-tasks and 38% for entire tasks – with only 30 demonstrations. This manifests the reliability of long-horizon data collection of our system to serve advanced manipulation task learning.

### B. Tactile Data as Policy Input

**Tactile Data Processing.** FSR sensors capture the pressure on the robot during object gripping. These signals can also be processed to calculate impulse by differentiating force over time. For tactile learning, touch signals are represented as vectors, encoded using MLPs, or visualized as point sets for visual embedding learning, akin to DexPoint [55].

**Results and Discussion.** As shown in Tab. VI, utilizing tactile data does not necessarily enhance outcomes, potentially because vision alone suffice for these tasks. Tactile

TABLE V: **DP3 Results for Long-Horizon Tasks.** SR-subtask is subtask success rate.

| Task           | DP3 [40]   |      |
|----------------|------------|------|
|                | SR-subtask | SR   |
| Wipe glass     | 7/10       | 7/10 |
| Sweep floor    | 8/10       | 4/10 |
| Prepare coffee | 7/10       | 4/10 |

TABLE VI: **Ablation of Touch as DP3 Policy Input.** Success rates are reported for two bimanual tasks.

| Signal  | Representation | Clean pan   | Uncover & pour |
|---------|----------------|-------------|----------------|
| /       | /              | <b>8/10</b> | 7/10           |
| force   | vector         | 7/10        | 4/10           |
| impulse | vector         | <b>8/10</b> | 7/10           |
| force   | point cloud    | 7/10        | 6/10           |
| impulse | point cloud    | 7/10        | <b>8/10</b>    |

feedback only activates upon the contact between hands and objects, thus not aiding in the object’s identification or localization. Moreover, using force data delivers slightly inferior performance compared to impulse data, possibly due to the zero-point drift in tactile readings over time. This drift, likely caused by the deformation of viscous materials used in sensor embedding, can persistently alter force reading during tasks, leading to unstable performance. Impulse, by contrast, is more robust against such variations.

## VI. CONCLUSION AND LIMITATION

**Limitation.** Our system has two main limitations: (i) Vision Pro’s hand tracking is inaccurate when fingers are self-occluded, causing jerky control commands. This could be mitigated by fusing data from wearable devices with Vision Pro. (ii) Subtle changes in haptic feedback are difficult to sense when the robot touches an object lightly. In future work, we will consider using actuators with larger contact areas and better sensitivity, such as piezoelectric actuators, to address this problem.

**Conclusion.** In this paper, we introduce Bunny-VisionPro, a bimanual dexterous teleoperation system for real-time robotic control. Our system utilizes Vision Pro to track hand poses and convert them into control commands, addressing critical challenges including loop joint hand retargeting, collision avoidance, and singularity to ensure accurate, safe, and responsive operations. Moreover, we develop a haptic feedback device that delivers tactile sensations to the operator, thus enhancing control accuracy and fostering a more immersive experience.

## REFERENCES

- [1] K. Hauser, "Recognition, prediction, and planning for assisted teleoperation of freeform tasks," *Autonomous Robots*, vol. 35, pp. 241–254, 2013.
- [2] N. Zhong and K. Hauser, "Attentiveness map estimation for haptic teleoperation of mobile robot obstacle avoidance and approach," *IEEE Robotics and Automation Letters*, 2024.
- [3] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube," in *Proceedings of Robotics: Science and Systems (RSS)*, 2022.
- [4] 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 *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- [5] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- [6] 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.
- [7] 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.
- [8] A. Toedtheide, X. Chen, H. Sadeghian, A. Naceri, and S. Haddadin, "A force-sensitive exoskeleton for teleoperation: An application in elderly care robotics," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 12 624–12 630.
- [9] 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.
- [10] A. Tung, J. Wong, A. Mandlekar, R. Martín-Martín, Y. Zhu, L. Fei-Fei, and S. Savarese, "Learning multi-arm manipulation through collaborative teleoperation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 9212–9219.
- [11] W. Zhao, J. Chai, and Y.-Q. Xu, "Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data," in *Proceedings of the ACM SIGGRAPH/eurographics symposium on computer animation*, 2012, pp. 33–42.
- [12] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang, "Semi-supervised 3d hand-object poses estimation with interactions in time," in *CVPR*, 2021.
- [13] H. Zhang, Z. Zhao, Y. Yu, K. Gui, X. Sheng, and X. Zhu, "A feasibility study on an intuitive teleoperation system combining imu with semg sensors," in *Intelligent Robotics and Applications: 11th International Conference, ICIRA 2018, Newcastle, NSW, Australia, August 9–11, 2018, Proceedings, Part I 11*. Springer, 2018, pp. 465–474.
- [14] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *ICRA*, 2018.
- [15] J. DelPreto, J. I. Lipton, L. Sanneman, A. J. Fay, C. Fourie, C. Choi, and D. Rus, "Helping robots learn: A human-robot master-apprentice model using demonstrations via virtual reality teleoperation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 10 226–10 233.
- [16] Y. Pan, C. Chen, D. Li, Z. Zhao, and J. Hong, "Augmented reality-based robot teleoperation system using rgb-d imaging and attitude teaching device," *Robotics and Computer-Integrated Manufacturing*, vol. 71, p. 102167, 2021.
- [17] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "Open teach: A versatile teleoperation system for robotic manipulation," *arXiv preprint arXiv:2403.07870*, 2024.
- [18] 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.
- [19] H. Liu, Z. Zhang, X. Xie, Y. Zhu, Y. Liu, Y. Wang, and S.-C. Zhu, "High-fidelity grasping in virtual reality using a glove-based system," in *2019 international conference on robotics and automation (icra)*. IEEE, 2019, pp. 5180–5186.
- [20] M. Mosbach, K. Moraw, and S. Behnke, "Accelerating interactive human-like manipulation learning with GPU-based simulation and high-quality demonstrations," in *Humanoids*, 2022.
- [21] M. Schwarz, C. Lenz, A. Rochow, M. Schreiber, and S. Behnke, "Nimbro avatar: Interactive immersive telepresence with force-feedback telemanipulation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 5312–5319.
- [22] MANUS Meta, "Manus meta quantum mocap metagloves," <https://www.manus-meta.com/products/quantum-mocap-metagloves>.
- [23] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *ICRA*, 2020.
- [24] S. Li, N. Hendrich, H. Liang, P. Ruppel, C. Zhang, and J. Zhang, "A dexterous hand-arm teleoperation system based on hand pose estimation and active vision," *IEEE Transactions on Cybernetics*, 2022.
- [25] Y. Qin, H. Su, and X. Wang, "From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation," *RA-L*, vol. 7, no. 4, pp. 10 873–10 881, 2022.
- [26] S. P. Arunachalam, I. Güzey, S. Chintala, and L. Pinto, "Holo-dex: Teaching dexterity with immersive mixed reality," *arXiv preprint arXiv:2210.06463*, 2022.
- [27] L. S. Yim, Q. T. Vo, C.-I. Huang, C.-R. Wang, W. McQueary, H.-C. Wang, H. Huang, and L.-F. Yu, "Wfh-vr: Teleoperating a robot arm to set a dining table across the globe via virtual reality," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4927–4934.
- [28] P. Ponomareva, D. Trinitatova, A. Fedoseev, I. Kalinov, and D. Tsetserukou, "Grasplook: a vr-based telemanipulation system with r-cnn-driven augmentation of virtual environment," in *2021 20th International Conference on Advanced Robotics (ICAR)*. IEEE, 2021, pp. 166–171.
- [29] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," *arXiv preprint arXiv:2404.16823*, 2024.
- [30] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.
- [31] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.
- [32] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choremanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.
- [33] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-maroon, M. Giménez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas, "A generalist agent," *Transactions on Machine Learning Research*, 2022, featured Certification, Outstanding Certification. [Online]. Available: <https://openreview.net/forum?id=likK0kHjvj>
- [34] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- [35] J. Pari, N. M. Shafullah, S. P. Arunachalam, and L. Pinto, "The surprising effectiveness of representation learning for visual imitation," in *Proceedings of Robotics: Science and Systems (RSS)*, 2022.
- [36] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," *arXiv preprint arXiv:2108.03298*, 2021.
- [37] S. Haldar, J. Pari, A. Rai, and L. Pinto, "Teach a robot to fish: Versatile imitation from one minute of demonstrations," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- [38] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu,

- and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," in *Proceedings of the 7th Conference on Robot Learning (CoRL)*, 2023.
- [39] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.
- [40] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.
- [41] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3d object manipulation," *arXiv:2306.14896*, 2023.
- [42] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki, "Act3d: 3d feature field transformers for multi-task robotic manipulation," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: <https://openreview.net/forum?id=HFJuX1uqs>
- [43] J. Grannen, Y. Wu, B. Vu, and D. Sadigh, "Stabilize to act: Learning to coordinate for bimanual manipulation," in *Conference on Robotic Learning (CoRL)*, 2023.
- [44] W. Yang, A. Anglaurd, R. S. Pieters, J. Pajarinen, and J.-K. Kämäräinen, "Seq2seq imitation learning for tactile feedback-based manipulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5829–5836.
- [45] Y. Park, "Teleoperation system using apple vision pro." [Online]. Available: <https://github.com/Improbable-AI/VisionProTeleop>
- [46] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang, "Dexmv: Imitation learning for dexterous manipulation from human videos," in *European Conference on Computer Vision*. Springer, 2022, pp. 570–587.
- [47] P. T. Boggs and J. W. Tolle, "Sequential quadratic programming," *Acta numerica*, vol. 4, pp. 1–51, 1995.
- [48] S. G. Johnson and J. Schueller, "Nlopt: Nonlinear optimization library," *Astrophysics Source Code Library*, pp. ascl–2111, 2021.
- [49] A. Dietrich, T. Wimbock, A. Albu-Schaffer, and G. Hirzinger, "Reactive whole-body control: Dynamic mobile manipulation using a large number of actuated degrees of freedom," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 20–33, 2012.
- [50] M. Bhardwaj, B. Sundaralingam, A. Mousavian, N. D. Ratliff, D. Fox, F. Ramos, and B. Boots, "Storm: An integrated framework for fast joint-space model-predictive control for reactive manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 750–759.
- [51] T. Yoshikawa, "Manipulability of robotic mechanisms," *The international journal of Robotics Research*, vol. 4, no. 2, pp. 3–9, 1985.
- [52] E. G. Gilbert, D. W. Johnson, and S. S. Keerthi, "A fast procedure for computing the distance between complex objects in three-dimensional space," *IEEE Journal on Robotics and Automation*, vol. 4, no. 2, pp. 193–203, 1988.
- [53] A. Barnea, C. Oprisan, and D. Olaru, "Force sensitive resistors calibration for use in gripping devices," *Mechanical testing and diagnosis*, vol. 2, no. 3, pp. 18–27, 2012.
- [54] G. S. Lakshmi and R. Hemavathi, "Calibration of force sensitive resistor used in force controlled grippers," *Int. J. Eng. Res. Technol.*, vol. 9, pp. 356–358, 2020.
- [55] Y. Qin, B. Huang, Z.-H. Yin, H. Su, and X. Wang, "Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 594–605.