Learning Visuotactile Skills with Two Multifingered Hands

Toru Lin, Yu Zhang*, Qiyang Li*, Haozhi Qi*, Brent Yi, Sergey Levine, and Jitendra Malik

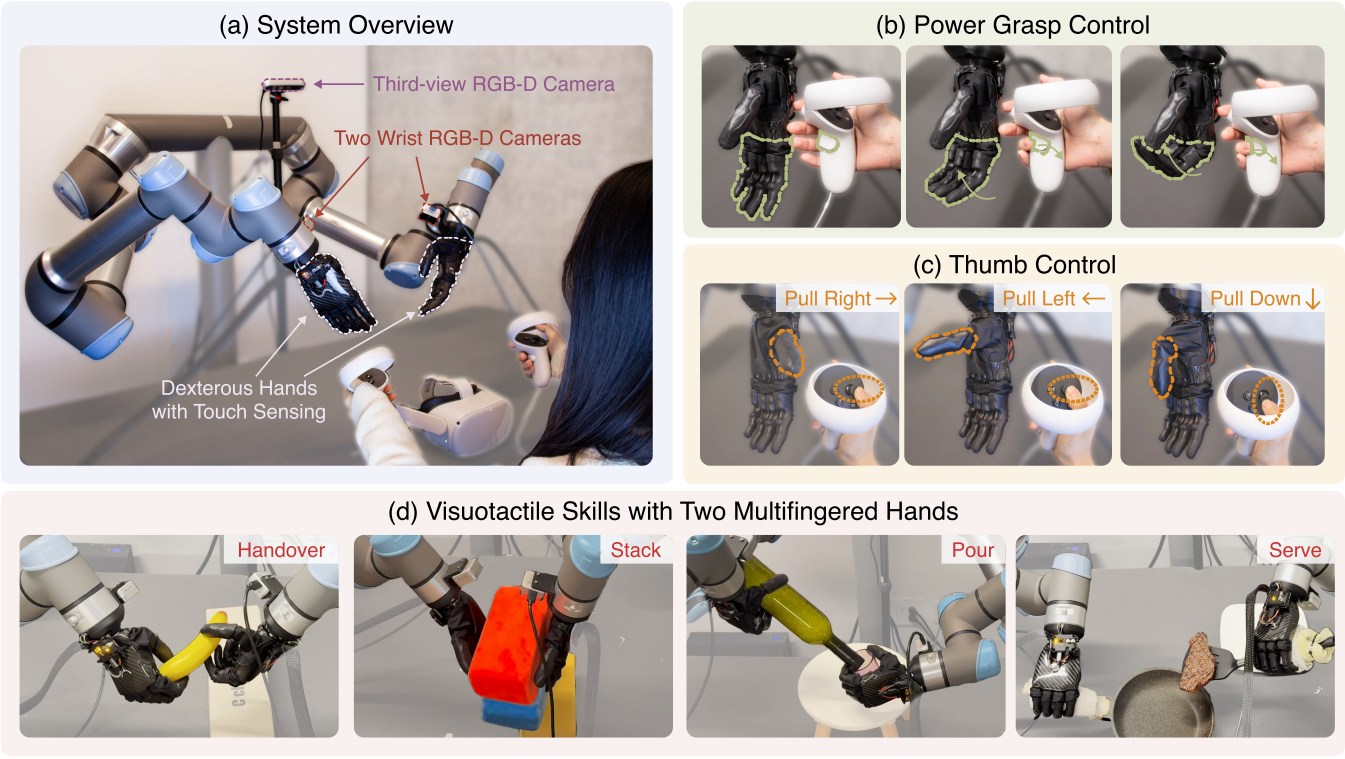

Fig. 1. **An overview of our system setup and learned visuotactile skills on four tasks.** (a) Our hardware and teleoperation system setup. The hardware consists of two UR5e robot arms, each equipped with a Psyonic Ability Hand. Visual observations are obtained via two wrist RGB-D cameras and one third-view RGB-D camera. Tactile observations come from the multifingered hands, with each fingertip equipped with six touch sensors. We utilize the Meta Quest 2 platform for teleoperation. (b) We use grip buttons of the Quest controllers to command power grasp of the non-thumb fingers. (c) We use thumbsticks to control the 2-DoF joint positions of the thumbs. (d) Four policies learned from visuotactile data collected by our hands-arms teleoperation system (HATO). These policies can accomplish a variety of complex bimanual tasks: handing over slippery object, stacking block tower, pouring from a wine bottle, and serving steak.

I. Introduction

Aiming to replicate human-like dexterity, perceptual experiences, and motion patterns, we explore learning from human demonstrations using a bimanual system with multifingered hands and visuotactile data [1], [2], [3]. Two significant challenges exist: the lack of an affordable and accessible teleoperation system suitable for a dual-arm setup with multifingered hands [4], and the scarcity of multifingered hand hardware equipped with touch sensing [5], [6]. To tackle the first challenge, we develop HATO, a low-cost hands-arms teleoperation system that leverages off-the-shelf electronics, complemented with a software suite that enables efficient data collection; the comprehensive software suite also supports multimodal data processing, scalable policy learning, and smooth policy deployment. To tackle the latter challenge, we introduce a novel hardware adaptation by repurposing two prosthetic hands equipped with touch sensors for research. Using visuotactile data collected from our system, we learn skills to complete long-horizon, high-precision tasks which are difficult to achieve without multifingered dexterity and touch feedback. Furthermore, we empirically investigate the effects of dataset size and sensing modality on policy learning. Our results mark a promising step forward in bimanual multifingered manipulation from visuotactile data.

* Equal contribution.
All authors are with University of California, Berkeley. Correspondence to toru@berkeley.edu

II. HATO: Hands-Arms Tele-Operation

We develop HATO, a novel teleoperation system for bimanual multifingered hands. Our system is easy to set up and intuitive to use, enabling efficient collection of bimanual dexterous manipulation data. An overview of our system is shown in Figure 1. For teleoperation of each hand-arm pair, HATO maps a Meta Quest 2 virtual reality (VR) controller's pose to the end-effector pose of the robot arm, and the controller's grip button and thumbstick to the hand's joint positions. The HATO software suite includes a data collection

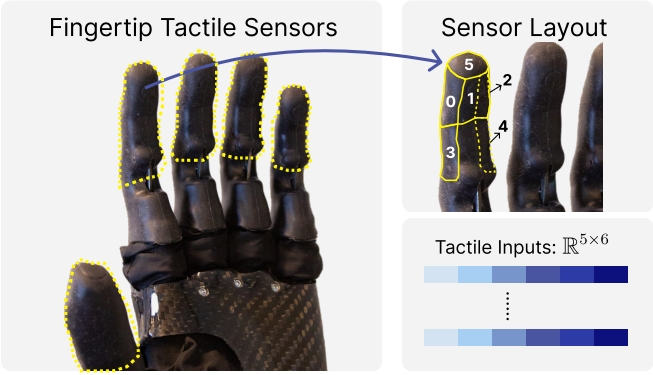

Fig. 2. **Fingertip Tactile Sensor Layout.** There are six tactile sensors on each of the fingertips. Each tactile sensor provides a continuous value proportional to the sensed pressure.

pipeline that records and processes data from all available sensing modalities (vision, touch, and proprioception).

For robot setup, we use two UR5e robot arms and attach two Psyonic Ability Hands as end effectors. These hands were originally designed for prosthetic use [7]; we repurpose them for research by designing custom printed circuit boards (PCBs) that simplify electrical wiring by integrating communication interfaces with power distribution. Each hand has five fingers and each finger has six actuated DoFs (one for each finger, two for the thumb). Each fingertip also comes with six touch sensors (see Figure 2).

Our teleoperation system leverages the Meta Quest 2 platform. It comes with a VR headset and a pair of controllers, each designated for one hand. Using a VR application like *oculus_reader* [8], one can stream data related to the controllers' poses and button states in real-time. Our main contribution is the development of a software suite that provides flexible options for translating movements detected by the Quest controllers to precise control commands for a bimanual multifingered robotic system. For arm control, we read the pose measurements from the Quest controller, and transform the pose to a desired end-effector (EEF) pose of the robot's coordinate system. For hand control, we map the controller's grip button to the joint positions of the four non-thumb fingers (4 DoF), and maps the thumbstick readings to joint positions of the thumb (2 DoF).

We collect multimodal data from both hands and arms by running HATO data collection pipeline at 10Hz. The data include the proprioceptive states of both the UR5e arms and the Ability Hands, the RGB-D images from three RealSense depth cameras (two mounted on the hand wrists, one mounted at a stationary "head-view" position), the touch sensor readings from the Ability hands, and the control commands given to the UR5e arms and the Ability hands. With visuotactile demonstration data collected from HATO, we learn a variety of bimanual dexterous skills for complex tasks using diffusion policies [9].

III. Experiments

We consider four challenging real-world tasks (Figure I) to study the bimanual dexterity enabled by our system.

We first qualitatively investigate whether having multi-fingered hands as end-effectors allows for better manipulation capabilities than parallel-jaw grippers by comparing their performances on the four manipulation tasks above. With multifingered hand end effectors, previously inexperienced teleoperators are able to collect hundreds of high-quality demonstrations within a few hours.

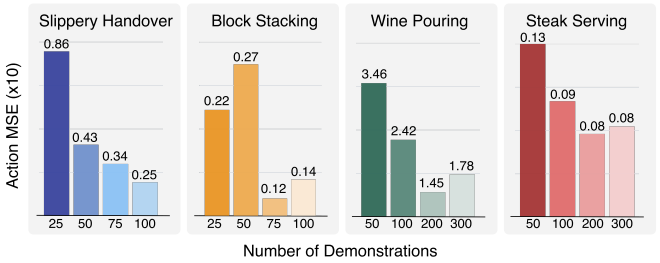

Fig. 3. **How does the demonstration dataset size affect policy prediction error?** Across all tasks, having more demonstration trajectories consistently lead to lower prediction loss. In particular, the policy performance saturates for *Bloock Stacking* at 75 demonstrations, *Wine Pouring* at 200 demonstrations and *Steak Serving* at 100 demonstrations.

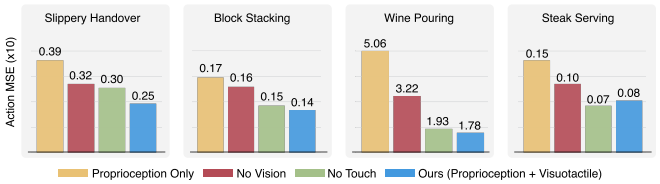

Fig. 4. **How vision and touch affect the policy performance across four challenging bimanual manipulation tasks.** Across all tasks, vision is crucial for the policy to achieve low prediction error.

Task	Handover	Stacking	Pouring	Serving
Pickup	10 / 10	10 / 10	10 / 10	10 / 10
Task Success	10 / 10	10 / 10	9 / 10	5 / 10

TABLE I. **Success rate on each of the four challenging bimanual manipulation tasks.** For *Slippery Handover* and *Wine Pouring*, we use only image observation and proprioceptive state as we find these two inputs are sufficient to achieve almost 100% success rate. For *Block Stacking* and *Steak Serving*, we use image, proprioception, and touch as input. The pickup success is an intermediate metric that measures how often the hands successfully pick up both objects.

We then validate the effectiveness of HATO as a data collection pipeline by demonstrating successful policies trained from HATO-collected datasets. In particular, we record the task success rate of learned policies using 10 deployment trails. In addition to the success rate for the full task, we also record how many times each policy successfully picks up the object(s) (e.g., bottle and cup for pouring, pan and spatula for steak serving, two blocks for stacking, and banana for handover) as the partial task completion rate. As shown in Table I, our policy is able to pick up the object(s) with 100% success rate across all tasks.

Finally, we study the efficiency of our learning method by empirically evaluating the correlation between number of demonstrations and policy performance (Figure 3), and quantitatively confirm that how the visuotactile sensing modalities are important to policy learning and performance (Figure 4).

REFERENCES

[1] F. Krebs and T. Asfour, "A bimanual manipulation taxonomy," *RA-L*, 2022.

[2] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile sensing—from humans to humanoids," *IEEE transactions on robotics*, vol. 26, no. 1, pp. 1–20, 2009.

[3] Z. Kappassov, J.-A. Corrales, and V. Perdereau, "Tactile sensing in dexterous robot hands," *Robotics and Autonomous Systems*, vol. 74, pp. 195–220, 2015.

[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 *RSS*, 2023.

[5] https://www.shadowrobot.com/dexterous-hand-series/.

[6] https://www.allegrohand.com/.

[7] A. Akhtar, J. A. Austin, J. M. Cornman, D. M. Bala, and Z. Wang, "System and method for an advanced prosthetic hand," Mar 2021.

[8] https://github.com/rail-berkeley/oculus_reader.

[9] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *RSS*, 2023.