Haptic-ACT: Bridging Human Intuition with Compliant Robotic Manipulation via Immersive VR

Kelin Li, Shubham M Wagh, Nitish Sharma, Saksham Bhadani, Wei Chen,
Chang Liu, and Petar Kormushev

***Abstract*— Robotic manipulation is essential for the widespread adoption of robots in industrial and home settings and has long been a focus within the robotics community. Advances in artificial intelligence have introduced promising learning-based methods to address this challenge, with imitation learning emerging as particularly effective. However, efficiently acquiring high-quality demonstrations remains a challenge. In this work, we introduce an immersive VR-based teleoperation setup designed to collect demonstrations from a remote human user. We also propose an imitation learning framework called Haptic Action Chunking with Transformers (Haptic-ACT). To evaluate the platform, we conducted a pick-and-place task and collected 50 demonstration episodes. Results indicate that the immersive VR platform significantly reduces demonstrator fingertip forces compared to systems without haptic feedback, enabling more delicate manipulation. Additionally, evaluations of the Haptic-ACT framework in both the MuJoCo simulator and on a real robot demonstrate its effectiveness in teaching robots more compliant manipulation compared to the original ACT. Additional materials are available at https://sites.google.com/view/hapticact.**

I. Introduction

With the growing demand for robotics to assist humans in daily manipulation tasks, robotic manipulation has garnered increasing attention from the robotics community. Over the past decades, it has made tremendous progress [1]–[6]. In these studies, robotic manipulation is typically performed by a robot arm equipped with a gripper attached to its end-effector. RGB-D cameras are commonly used to observe the environment and capture visual information, including the poses and geometric features of objects. These features can be represented as either 2D RGB images [7]–[9] or 3D point clouds [10], [11], valid manipulations will be generated based on the observed object features. With the development of learning-based methods, efficiency and generalizability have become important considerations in the design of frameworks [4], [12], [13]. In recent years, language models have been integrated with visual models to enable robots to handle a wide range of environments [6], [7], [14].

Although existing robotic manipulation methods can produce stable actions, a gap remains in applying traditional robot learning methods to real-world setups. Traditional robot learning methods involve the robot exploring the entire manipulation space to find a solution for a specific task. However, this process is usually inefficient and time-consuming, as it often involves redundant learning before arriving at an optimal solution. An efficient alternative to training robots in manipulation tasks is imitation learning, also known as Learning from Demonstration (LfD). In this approach, the robot learns by observing expert demonstrations, allowing skills to generalize to unseen scenarios. This process not only extracts information about the expert's behavior and the environment but also learns the mapping between observations and actions. [15]. Thus, the robot can learn in the correct direction to perform manipulation tasks effectively. In recent years, imitation learning has been extensively studied for enabling robots to perform various manipulation tasks [16]–[18]. However, efficiently collecting demonstrations using an appropriate platform remains a challenge.

To address this issue, this work introduces an immersive VR-based setup for teleoperation to collect demonstrations from human demonstrators. Additionally, an imitation learning framework called Haptic-ACT is proposed. The summary

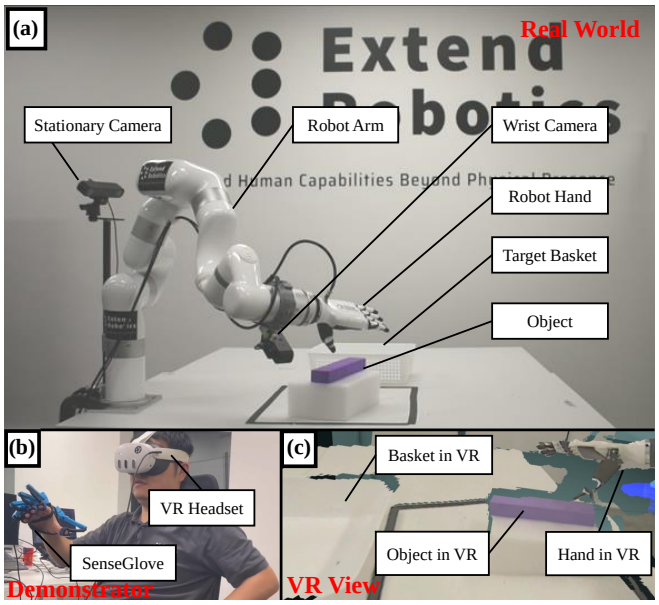

Fig. 1. Summary diagram of the proposed immersive VR-based setup used in this work, featuring a VR headset, a haptic feedback glove, a follower robot arm, and a robot hand. (a) illustrates the robot arm and hand system following human demonstrations and providing sensory feedback, (b) depicts the demonstrator remotely controlling the robot, and (c) displays the VR view from the headset.

Kelin Li is jointly with the Robot Intelligence Lab, Imperial College London, and the Extend Robotics, k.li20@imperial.ac.uk. Wei Chen, and Petar Kormushev are with the Robot Intelligence Lab, Imperial College London, 25 Exhibition Road, London, SW7 2DB, UK, (w.chen21, p.kormushev)@imperial.ac.uk. Shubham M Wagh, Nitish Sharma, Saksham Bhadani, and Chang Liu are with Extend Robotics, 5-9 Merchants Pl, Reading, RG1 1DT, UK (shubham.wagh, nitish, saksham.bhadani, chang.liu)@extendrobotics.com.

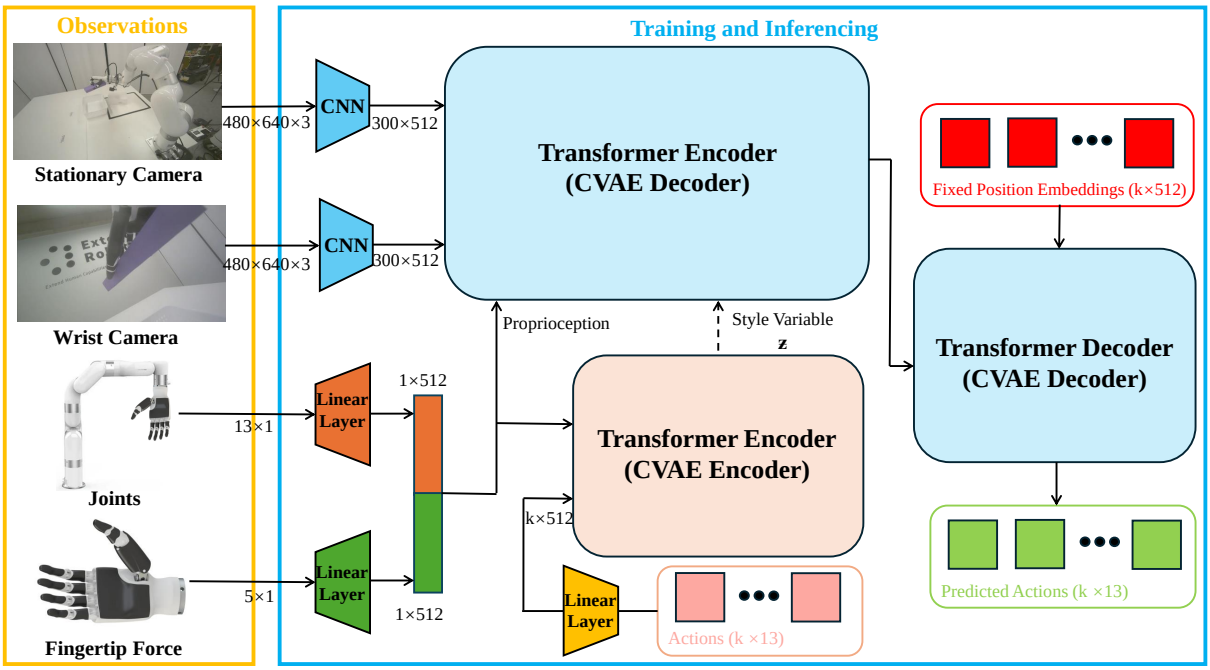

Fig. 2. Flowchart of the proposed Haptic-ACT. The observations include RGB images from two cameras, the robot's joint positions, and the fingertip forces of the hand. Note that the transformer encoder (CVAE encoder) operates only during the training phase to compute the style variable for the transformer encoder (CVAE decoder). During the inference phase, the style variable is fixed at 0.

diagram of the proposed immersive VR-based setup is shown in Fig. 1. The robot system consists of an xArm7 robot arm equipped with an Inspire robot hand and two ZED cameras. The human demonstrator remotely teleoperates the robot using a Meta Quest 3 headset and a SenseGlove, which tracks the hand's movements and maps them to the robot's joint positions using inverse kinematics (IK). The camera feed is rendered in the VR headset, enabling the demonstrator to see the robot's perspective in real-time. The fingertip contact force from the robot is mapped to the SenseGlove motor torque, allowing the demonstrator to experience an immersive demonstration [19]. The framework of the proposed Haptic-ACT is shown in Fig. 2. The observations include RGB images from the cameras, the robot's joint positions, and fingertip contact forces. The haptic information enables the robot to learn how to make soft contact with objects.

We summarize our main contributions as follows: (1) A VR-based setup that allows human demonstrators to teleoperate robots immersively. (2) The integration of SenseGlove, which provides haptic feedback to enhance teleoperation. (3) The proposed Haptic-ACT, which enables robots to learn more compliant manipulations compared to the original ACT.

II. Related Work

A. Robotic Manipulation

Robotic manipulation, encompassing tasks such as grasping, moving, and reorienting objects, is a fundamental capability in robotics. These tasks often require varying levels of contact with the environment, making precise control of contact forces—whether implicitly or explicitly—crucial for successful execution. As robots increasingly take on roles traditionally performed by humans, research on robotic manipulation has expanded significantly [20], [21].

Manipulation tasks frequently involve contact-rich interactions, such as grasping a hammer for hammering [22], screwing on a bottle cap [3], or folding clothes [2], [23]. The most intuitive approach to robotic manipulation is to design controllers based on control theory. Among various control strategies, impedance control is particularly notable for enabling desired dynamic interactions between a manipulator and its environment. This method regulates the dynamic relationship between the manipulator's motion variables and the contact forces, making it widely used for force tracking [24], human-robot interaction [25], and other applications. However, traditional controllers often struggle with adaptability, handling only a limited range of tasks and failing in unforeseen situations.

In recent years, learning-based approaches have gained prominence in robotic manipulation [4], [8], [10]. Among these, imitation learning has demonstrated the greatest potential for improving manipulation performance [18], [26]. A key advantage of imitation learning is its ability to leverage human expertise to teach robots complex manipulation skills without requiring explicit programming. Techniques such as behavior cloning and inverse reinforcement learning enable robots to generalize from expert demonstrations and adapt to various tasks [27], [28]. However, effectively acquiring high-quality demonstrations remains a significant challenge.

B. Platforms for Demonstrations

As discussed earlier, effectively gathering demonstrations with an appropriate platform is crucial for acquiring high-