EasyReach: Interactive Preference Manager for Vision-Language-Action Models

Yu-Wei Chang
UCLA
ywchang@ucla.edu

Leyi Zou
UCLA
zelozou@ucla.edu

Abstract

Vision-Language-Action models excel at generalist robotic manipulation but regress to the mean of their training data, ignoring user-specific preferences such as handedness, interaction speed, and safety margins. We introduce EasyReach, a system-level approach that enables zero-shot personalization through an "Interaction Manager" framework. Using a dual-camera setup—wrist-mounted for task context and external for human monitoring—a lightweight VLM observes user signals (verbal feedback, posture, proximity) and synthesizes natural language prompts that guide a frozen VLA in real-time, without retraining or visual in-context learning latency.

Introduction

Vision-Language-Action models like RT-2[1] and OpenVLA[2] generalize across diverse scenes but regress to the mean of their training data, failing to personalize for user-specific preferences (handedness, reaction speed, proximity comfort). Fine-tuning per user causes catastrophic forgetting, while visual in-context learning incurs prohibitive latency.

We observe that VLAs already follow nuanced natural language instructions—the problem is system perception. Standard deployments are "blind" to human context: the wrist camera sees the task, not the user's position or body language.

EasyReach addresses this through a dual-camera architecture. A wrist-mounted camera provides task context to the frozen VLA, while an external camera feeds an "Interaction Manager"—a lightweight VLM that observes human-robot spatial dynamics and translates user signals (posture, verbal feedback, proximity) into real-time prompts. This preserves VLA generalization while enabling zero-shot personalization through natural language conditioning.

Methods

System Architecture. We deploy on a AUBO 6-DOF manipulator with a dual-camera setup: (1) wrist-mounted RGB camera provides task context to the frozen VLA; (2) external camera monitors human-robot spatial relationships. A directional microphone captures verbal feedback.

Control Pipeline. The wrist camera provides RGB to the frozen VLA. The external camera feeds an Interaction Manager (lightweight VLM: GPT-4o-mini or Gemini Flash) that extracts user signals: handedness, proximity violations (< 20cm), postural recoil, and anthropometric features (height, reach). The Manager performs zero-shot visual reasoning and synthesizes natural language prompts that modulate VLA behavior.

Example prompt: *"The user is approaching with their LEFT hand. They are tall (1.8m), so maintain a high approach angle. CRITICAL: The user just leaned back—reduce velocity to 30% and retract 10cm."*

The VLA receives the wrist image and synthesized prompt, then outputs 6-DOF end-effector commands executed by the robot controller.

Evaluation Goals

We evaluate EasyReach on a physical AUBO arm with unseen users across four dimensions: (1) **Zero-shot personalization** measuring handover success rates for left/right handedness and height-adjusted approach angles; (2) **Real-time safety** measuring prompt synthesis latency (< 100ms target) and response to proximity violations/postural recoil; (3) **VLM-based inference accuracy** validating that visual reasoning correctly identifies user preferences without prior data collection; (4) **Model compatibility** demonstrating integration with frozen VLA models (OpenVLA, RT-2) without modification.

Bibliography

[1] A. Brohan *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," *arXiv preprint arXiv:2307.15818*, 2023.

[2] M. J. Kim *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model," *arXiv preprint arXiv:2406.09246*, 2024.