PolyUMI: Visual + Auditory + Tactile Manipulation Platform for Imitation Learning

Anonymous Author(s)

***Abstract*— Contact-rich manipulation remains a fundamental challenge in robot learning, in part because contact events are brief, highly variable, and not fully captured by vision alone. We present PolyUMI, a real-time multi-modal data collection and control platform that unifies four sensing modalities in a single end-effector: optical tactile sensing, mechanical vibration (via contact microphone), egocentric vision, and proprioception. Building on the Universal Manipulation Interface (UMI) [1] handheld gripper framework, PolyUMI adds a custom touch-sensing finger inspired by PolyTouch [2], delivering synchronized streams of tactile video, contact audio, wrist camera video, and pose data—all from a fully wireless, battery-powered gripper. The system also supports an end-effector for the Franka Panda arm that preserves the same sensor geometry as the handheld gripper to facilitate policy transfer. We describe the hardware, firmware, and software architecture of PolyUMI and discuss its potential as a platform for studying how tactile and auditory sensing can complement vision in learning contact-rich manipulation policies.**

I. Introduction

Learning dexterous manipulation from demonstration requires policies that can detect and respond to contact. However, most imitation learning platforms treat manipulation as a purely visual-proprioceptive problem, discarding the rich tactile and auditory signals that humans exploit when handling objects—detecting a grasp slip, registering a surface texture, or hearing the moment a part seats into a socket.

Efforts to close this gap have explored tactile sensors [3], [4], [5], contact microphones [6], and their combination [2] as complementary inputs to vision-based robot learning. However, integrating these modalities into a cohesive, hardware-software co-designed platform for scalable demonstration collection remains an open engineering challenge. Existing platforms either support limited sensor combinations, or constrain data collection to tethered setups/teleoperation.

We present **PolyUMI** (**Poly**modal **U**niversal **M**anipulation **I**nterface), which addresses this gap by combining:

- A custom optical tactile finger with an integrated contact microphone;
- A GoPro wrist camera with fisheye optics and side mirrors, following the UMI design [1];
- Embodiment-agnostic proprioception via monocular-inertial SLAM (handheld gripper) or forward kinematics (robot arm);
- A fully wireless, single-button data collection workflow with time-synchronized outputs ready for policy training.

Project page with videos: `https://cwoodhayes.github.io/projects/polyumi/`

Furthermore, PolyUMI is 100% open source and is designed to be comparatively cheap and simple for anyone to build, especially compared to existing touch sensing systems; the entire electro-mechanical system can be made using a 3D printer, a soldering station, and parts available on Amazon and DigiKey, and the sensing surface (based off of PolyTouch [2]) can be manufactured in 15 minutes with no prior experience, since it uses VHB tape for its comformable elastomer rather than a gel.

PolyUMI's software is built from the ground up on a modern robotics stack (ROS 2, Python 3.13, and Foxglove for visualization) and the system is designed to support rapid hardware-software co-iteration—from changing the sensing surface material to swapping in a new end-effector—without changing the data pipeline or making significant changes to the mechanical design.

II. System Overview

A. Hardware

PolyUMI consists of two complementary hardware forms (Fig. 1): a **handheld gripper** for in-the-wild demonstration collection, and a **Franka Panda end-effector** for robot-mounted inference. Both share the same finger design, enabling hardware reuse and maintaining consistent sensor geometry to support policy transfer.

The finger itself uses a modular multi-part design such that the sensing surface is removeable and replaceable in 30 seconds or less, while the finger itself can be detached from the gripper and end-effector in 5 minutes or less.

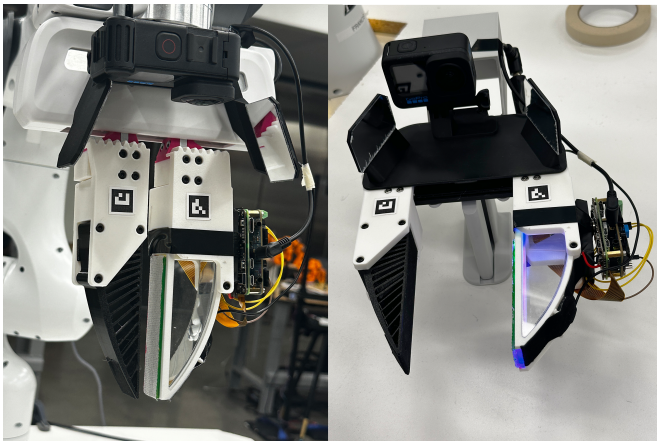


Fig. 1: PolyUMI handheld gripper (right) and Franka Panda end-effector (left), sharing the same tactile finger design.

The gripper is fully wireless with approximately 5 hours of battery life, controlled entirely from an onboard Raspberry Pi Zero 2W. A single button press starts and stops synchronized recording across all sensor modalities. The Franka end-effector uses a finger-to-mount adapter that allows the same physical sensing finger to transfer between the handheld gripper and the robot arm—which we expect will reduce the domain gap for training and improve policy performance.

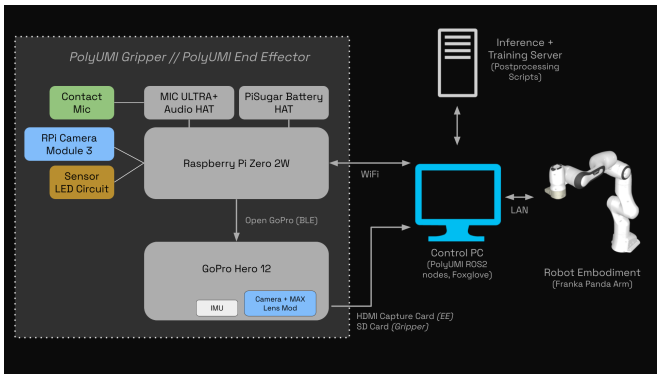


Fig. 2: System block diagram. All sensing modalities are time-synchronized to a single clock on the onboard Raspberry Pi.

B. Sensor Modalities

Table I summarizes the four sensor modalities and their output specifications.

TABLE I: PolyUMI Sensor Modalities

Modality	Sensor	Output
Tactile	Optical finger (custom)	10 fps, 540×480 MJPEG
Vibration	Contact mic (custom mount)	16 kHz mono PCM
Vision	GoPro Hero 12 + fisheye	60 fps, 1920×1080 MP4
Proprioception	SLAM / FK / ArUco	6-DoF pose + gripper width

Optical tactile finger.

The sensing finger follows the same core principle as PolyTouch [2]: a deformable gel-like surface is observed by an internal camera via a curved mirror, giving a near-overhead view of the contact area (Fig. 3). The sensing surface uses VHB tape coated with aluminum powder and covered with medical tape for reflectivity, durability, and texture. All hardware, firmware, and manufacturing processes were independently designed and implemented using the published PolyTouch description as a starting point; no source code or CAD from that work was used.

Contact microphone. A contact microphone is rigidly coupled to the finger housing, capturing mechanical vibrations propagating through the sensor body rather than airborne sound. This modality is well-suited for detecting the onset and character of contact events (hard tap vs. soft press, sliding vs. static) [6]. Audio is captured at 16 kHz via a Raspiaudio ULTRA+ HAT on the Raspberry Pi.

Wrist camera. The GoPro Hero 12 with MAX Lens Mod 2.0 (≈177° FOV) follows the UMI design (though we use a newer GoPro model to match current availability), with side mirrors providing a binocular view of the manipulated object. PolyUMI adds BLE-based GoPro control from the onboard Raspberry Pi to ensure synchronized recording start across all data streams.

Proprioception. Proprioception is handled differently depending on embodiment (Fig. 4). When using the handheld gripper, PolyUMI runs ORB-SLAM3 [7] monocular-inertial SLAM on the GoPro's video feed and IMU to derive a 6-DoF pose trajectory; gripper width is estimated from ArUco tags on the fingers. When mounted on the Franka arm, joint angles come directly from `libfranka` and EE pose from forward kinematics. Policies trained by the system can operate either in embodiment-specific joint space (using an IK solver like OMPL to generate joint angles from the UMI trajectory) or in embodiment-agnostic cartesian space (using FK to derive end-effector pose for the robot embodiment).

C. Data Pipeline

The data pipeline is designed to minimize friction from hardware to training-ready dataset. Key design priorities are: (1) *as close to turnkey as possible*—a clean hardware setup goes from zero to live-streaming in under 20 minutes; (2) *single-button recording*—all streams start and stop together; and (3) *training-ready output*—all data is timestamped, synchronized, and stored per-episode as MCAP files with comprehensive metadata. A postprocessing command fetches all recorded episodes from the gripper over the network. Pipeline extensions convert the at-rest MCAP format to common training formats (Zarr, LeRobot Dataset). All demonstrations can be replayed or live-streamed in Foxglove (Fig. 5).

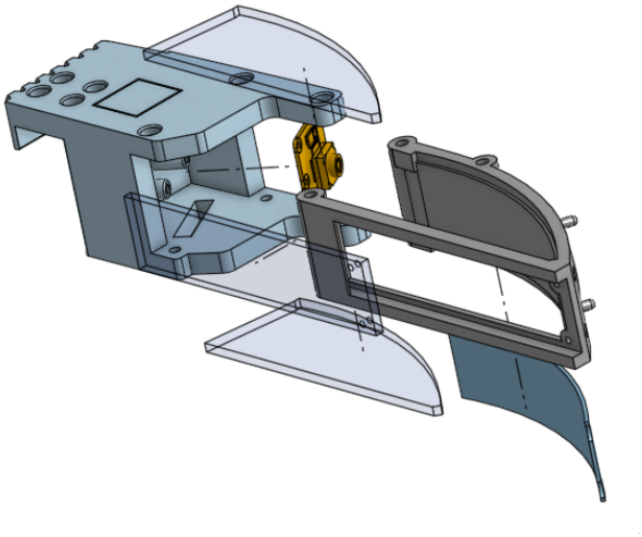
III. Discussion

Motivation: hardware-software co-design for contact-rich learning. A central motivation for this work is the belief that as learned models take over increasing portions of the robot control stack, hardware becomes a more fluid design variable. PolyUMI is explicitly designed to support iteration at the hardware level—new sensing surface materials, different mirror geometries, or alternative end-effector designs—without requiring changes to the data pipeline or training infrastructure. We see this as a meaningful shift in how one can approach the data-hardware-model loop for contact-rich manipulation.

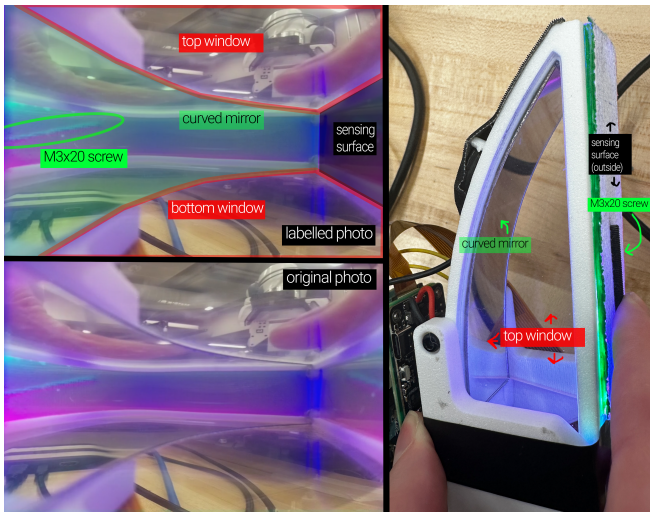
Relevance to this workshop. PolyUMI is positioned at the intersection of two themes of this workshop: *perception and high-dimensional sensing for contact-rich tasks*, and *tools and infrastructure for contact-rich robotics*. By combining optical tactile sensing, contact audio, and egocentric vision in a single portable platform, PolyUMI is designed to make it tractable to study how these complementary modalities interact in learned policies for contact-rich manipulation.

Open questions. Several open questions motivate continued development of this platform:

- *Which sensor modalities matter, and when?* Tactile and audio signals are most informative at contact events,



(a) Exploded CAD view of the tactile finger.



(b) Finger camera field of view (annotated).

Fig. 3: PolyUMI optical tactile finger. The curved mirror gives a consistent optical distance from lens to gel surface so the full contact area stays in-focus.

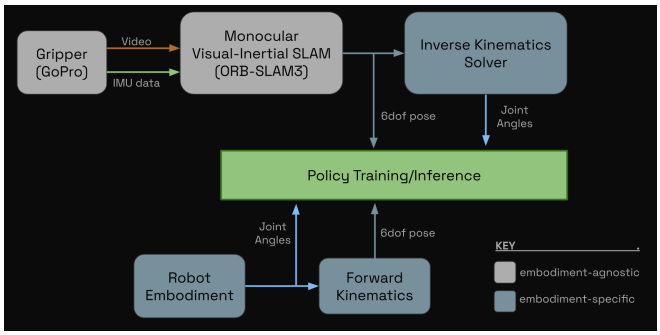


Fig. 4: Proprioception dataflow. An embodiment-agnostic 6-DoF pose + gripper width representation is produced regardless of whether data comes from the handheld gripper (via SLAM) or a robot arm (via FK).

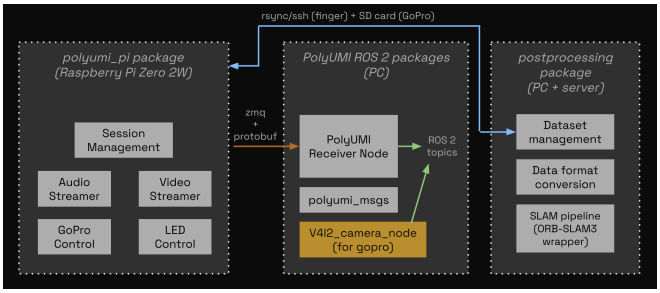


Fig. 5: PolyUMI software component diagram, spanning onboard firmware (Raspberry Pi), GoPro BLE control, post-processing, and visualization (Foxglove).

which are sparse in time. How should a policy architecture weight these modalities across different task phases?

- *Cross-embodiment transfer.* How much does the shared finger geometry between the handheld gripper and the Franka end-effector actually reduce the sim-to-real / demo-to-policy gap? Preliminary hardware design suggests this should help; controlled experiments are planned. Data augmentation strategies following ManiWAV [6] are planned to address the more significant domain gap in the audio modality, as the differing internal mechanisms for the arm and gripper create different noise profiles.
- *Sensing surface design.* The stiffness and reflectivity of the sensing surface significantly affect both tactile signal quality and the ability to generalize across objects. Systematic exploration of materials is an open hardware research question.

IV. Conclusion

PolyUMI is a multi-modal manipulation data collection and control platform that integrates optical tactile sensing, contact microphone audio, fisheye wrist vision, and embodiment-agnostic proprioception into a single wireless end-effector compatible with in-the-wild demonstration collection and robot arm deployment. We are currently developing training and inference infrastructure to evaluate imitation learning policies (Diffusion Policy, ACT, and vision-language-action models) on contact-rich tasks using this platform. We believe PolyUMI can serve as both a practical tool for multi-modal robot learning research and a template for hardware-software co-design in the contact-rich manipulation setting.

Hardware designs, firmware, and software are available at: `https://cwoodhayes.github.io/projects/`

polyumi/.

REFERENCES

[1] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[2] J. Zhao, N. Kuppuswamy, S. Feng, B. Burchfiel, and E. Adelson, "Polytouch: A robust multi-modal tactile sensor for contact-rich manipulation using tactile-diffusion policies," 2025. [Online]. Available: https://arxiv.org/abs/2504.19341

[3] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, 2017. [Online]. Available: https://www.mdpi.com/1424-8220/17/12/2762

[4] W. K. Do and M. K. III, "Densetact: Optical tactile sensor for dense shape reconstruction," 2022. [Online]. Available: https://arxiv.org/abs/2201.01367

[5] Y. Li, Y. Chen, Z. Zhao, P. Li, T. Liu, S. Huang, and Y. Zhu, "Simultaneous tactile-visual perception for learning multimodal robot manipulation," *IEEE Robotics and Automation Letters*, vol. 11, no. 4, pp. 5254–5261, 2026.

[6] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, "Maniwav: Learning robot manipulation from in-the-wild audio-visual data," *arXiv preprint arXiv:2406.19464*, 2024.

[7] C. Campos, R. Elvira, J. J. Gómez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.