PolyUMI

Visual + Auditory + Tactile Manipulation Platform for Imitation Learning

Conor Hayes (cwhayes@u.northwestern.edu)

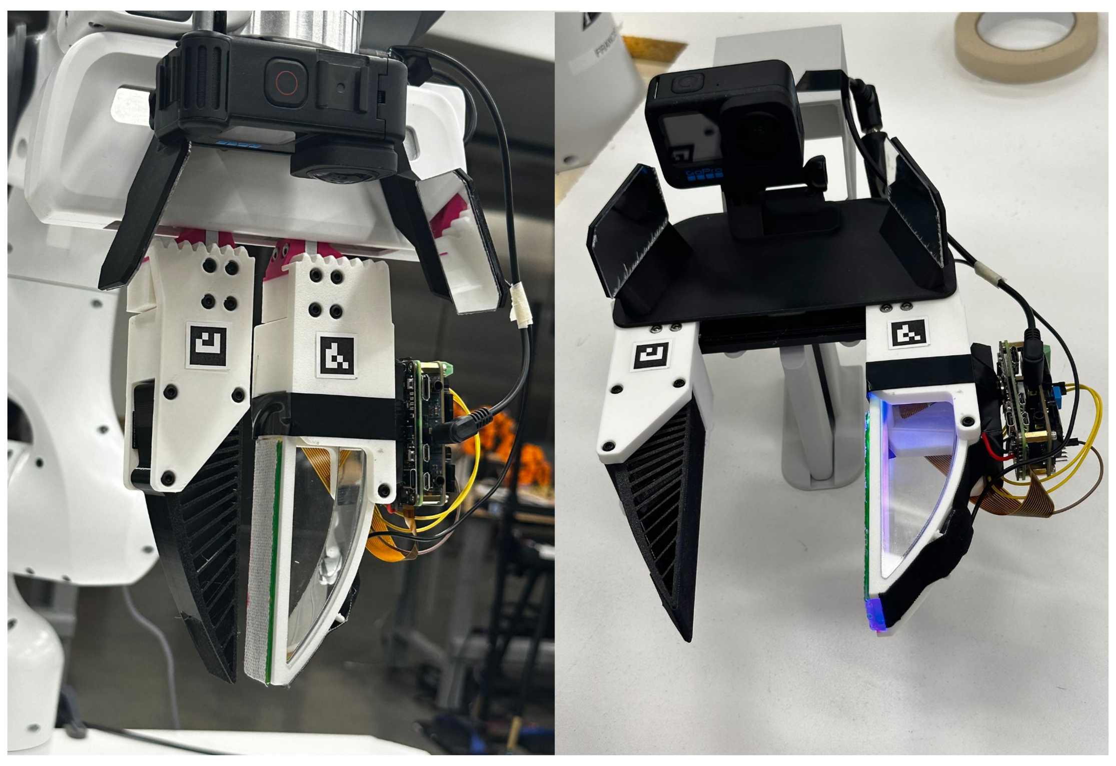
Center for Robotics and Biosystems, Northwestern University, Evanston, IL, USA

Overview: **PolyUMI is an integrated hardware and software system for training and deploying imitation learning policies. Building on the *Universal Manipulation Interface* [1] and *PolyTouch* [6], it is the first such system to enable collecting synchronized visual, auditory, tactile, and proprioceptive data from human demonstrations in-the-wild. We open source this platform for the community to use alongside our own research.**

Introduction

Project Aims:

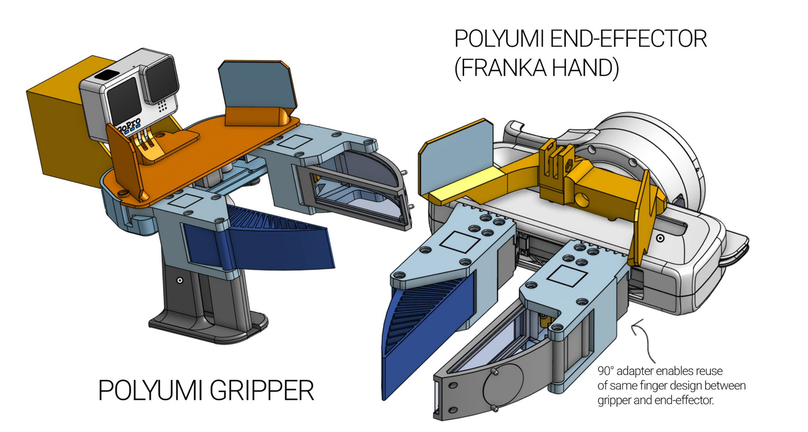
- Incorporate & evaluate **a novel set of sensor modalities for a UMI platform**, toward improved performance on dexterous + contact-rich manipulation
- **100% open source**, using commodity parts where possible for reduced BOM cost & ease of manufacture
- **Modular HW/SW design** to enable rapid prototyping & iteration across entire system

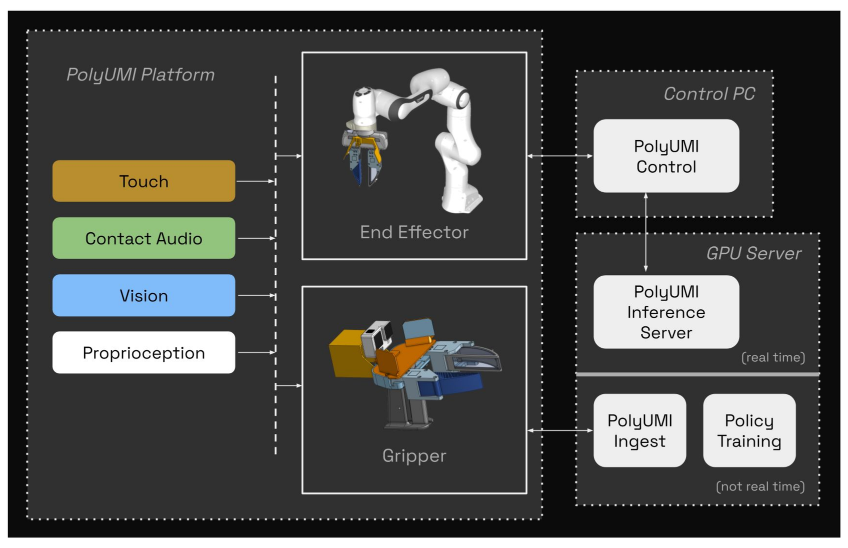


System	Vision	Touch	Vibration	Trains Anywhere	Open Source	In-house touch sensor
UMI [1]	✓			✓	✓	
Touch in the Wild [2]	✓	✓		∼	✓	✓
ManiWAV [3]	✓		✓	✓	✓	
TacThru [4]	✓	✓		∼	✓	✓
exUMI [5]	✓	✓		✓	✓	
ViTaMin [6]	✓	✓		✓		✓
PolyTouch [7]	✓	✓	✓			✓
PolyUMI	✓	✓	✓	✓	✓	✓

Comparison with similar systems: By adapting PolyTouch's [6] optical tactile + vibration sensor to function on a mobile gripper, as well as using tactics from UMI [1] and ManiWAV [3] to address resulting domain transfer issues, PolyUMI is able to capture natural, in-the-wild human demonstrations with a set of sensor modalities that have never before left the lab.
Note: [2] and [4]'s grippers require wired connections to operate and thus are slightly less flexible than those marked with a ✓

System Overview





PolyUMI Gripper: Handheld system for collecting human demonstrations in any environment. Fully wireless, with 5 hrs of battery life and no need for an external PC – recording starts & stops with the press of a button. Controlled by a Raspberry Pi Zero 2W.

PolyUMI End-Effector: Modular system for mounting the PolyUMI finger + other sensors to a robotic arm (currently the Franka Panda). Sensor layout and geometry is identical to the Gripper, to smooth domain transfer from gripper-based training to arm-based inference.

PolyUMI Control: ROS2-based system to ingest data streamed from the End-Effector and arm, make client requests to the Inference Server, and control the arm accordingly. Runs on NUC, user PC, or GPU server.

PolyUMI Inference Server *(work in progress)*: Exposes network-based access to the trained policy through an API endpoint; runs on a GPU server.

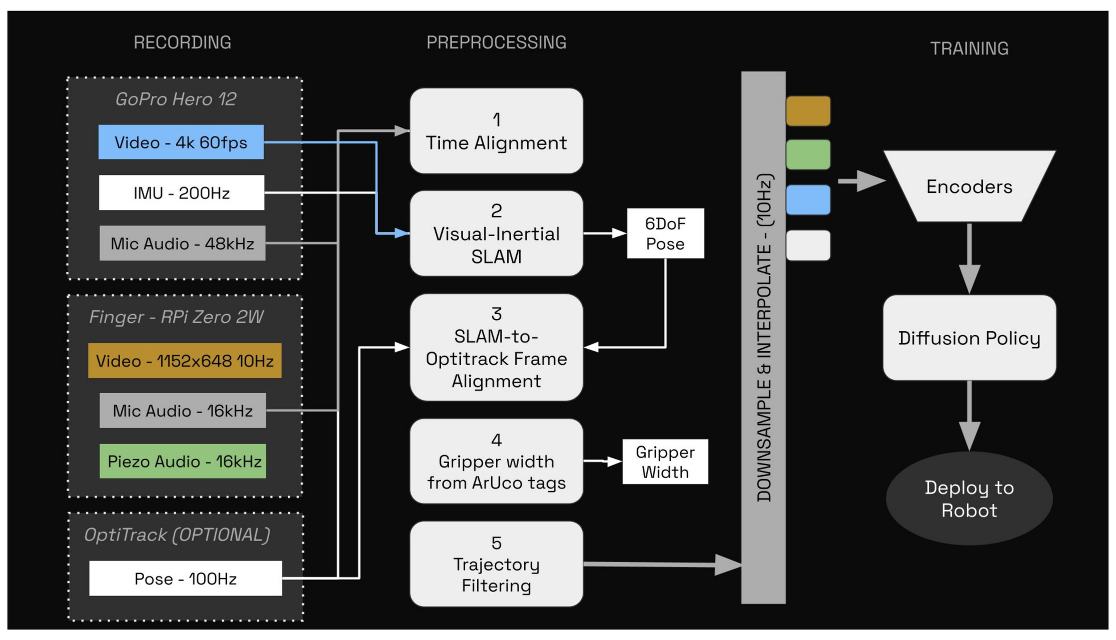
PolyUMI Ingest: Ingests demonstration data recorded on the PolyUMI gripper, pre-processes it (see below), and exports it in the appropriate model training format.

This separation of concerns between *PControl* and *PInference* allows us to leverage the ROS 2 ecosystem for visualization, middleware, and control, without interfering with the inference environment + server. In practice this greatly simplifies the deployment of this system, bringing it closer to turnkey software.

Data Collection + Preprocessing

Novel Contributions:

- Audio-based time alignment for precise synchronization between camera systems
- Use of OptiTrack for SLAM validation + training data for indoor scenes
- Possible: Kinematic trajectory filtering

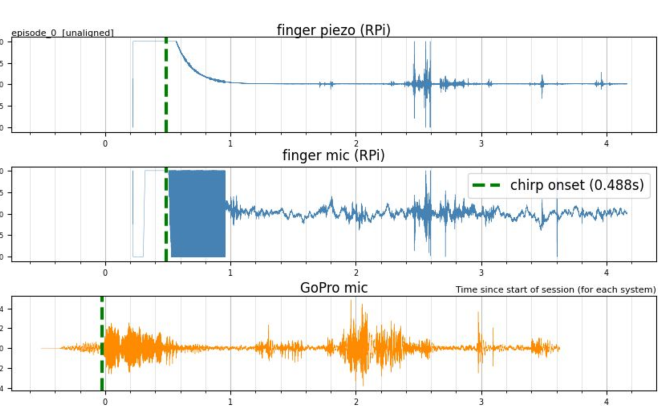


1 - Time Alignment

3 time domains to align: finger (touch + vibration), gopro (video + pose), and OptiTrack (pose).

OptiTrack is synced to the finger at start of recording via a pulse to a DIO pin on the Pi. GoPro is synced to the finger as follows:

- GoPro recording triggered by Pi over BLE; coarse accuracy of ~1s.
- At start of episode, Pi emits chirp using piezo buzzer (linear sweep over known frequency range).
- In post, matched filter to find chirp in both finger & gopro audio, giving precise time offset between the two domains.



2 - Visual-Inertial SLAM

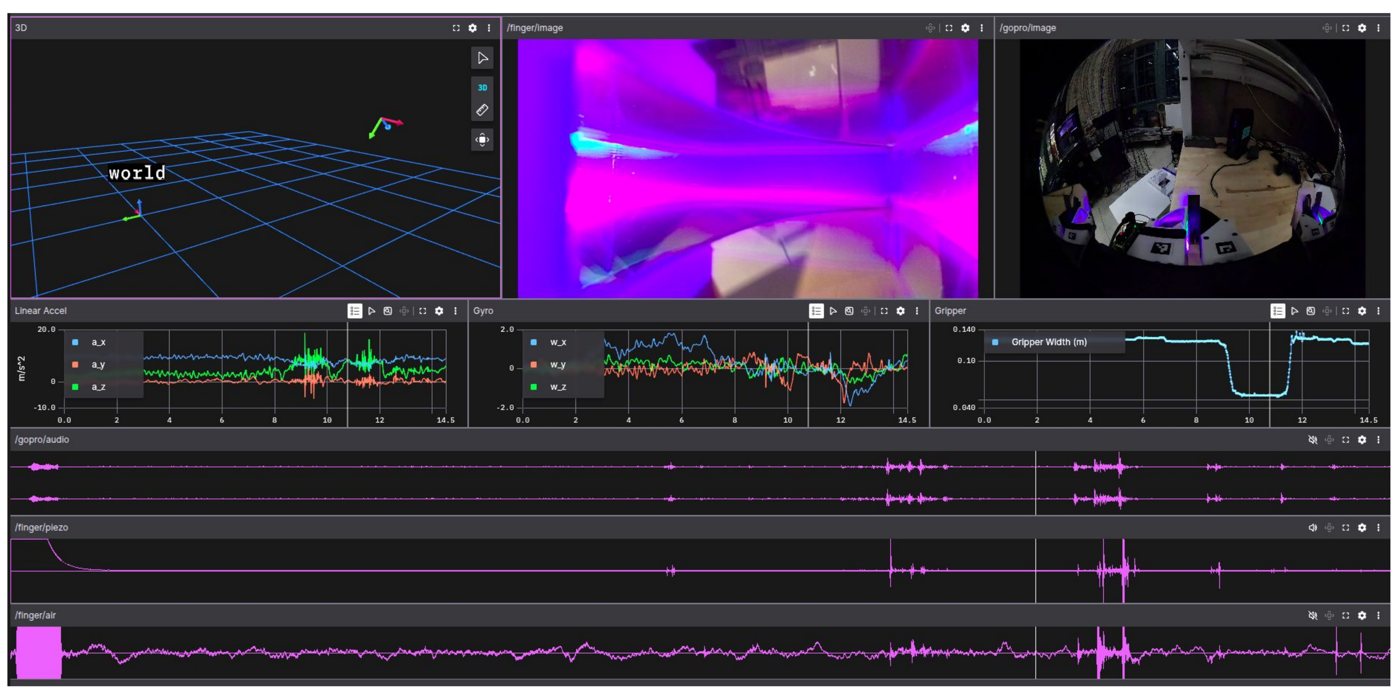
Following UMI, ORB-SLAM3 [8] is used for pose estimation in the wild from GoPro video + IMU data (see data visualizer at right). Minor fixes were made in a custom fork to accommodate updated camera hardware, fix crashes, and integrate with PolyUMI Ingest

3 - SLAM-to-Optitrack Alignment

To enable direct comparison of the SLAM and OptiTrack trajectories (if both are available), an SE(3) transform between the two frames is derived using Horn's method [10] for least-squares optimal alignment between two paired point sets.

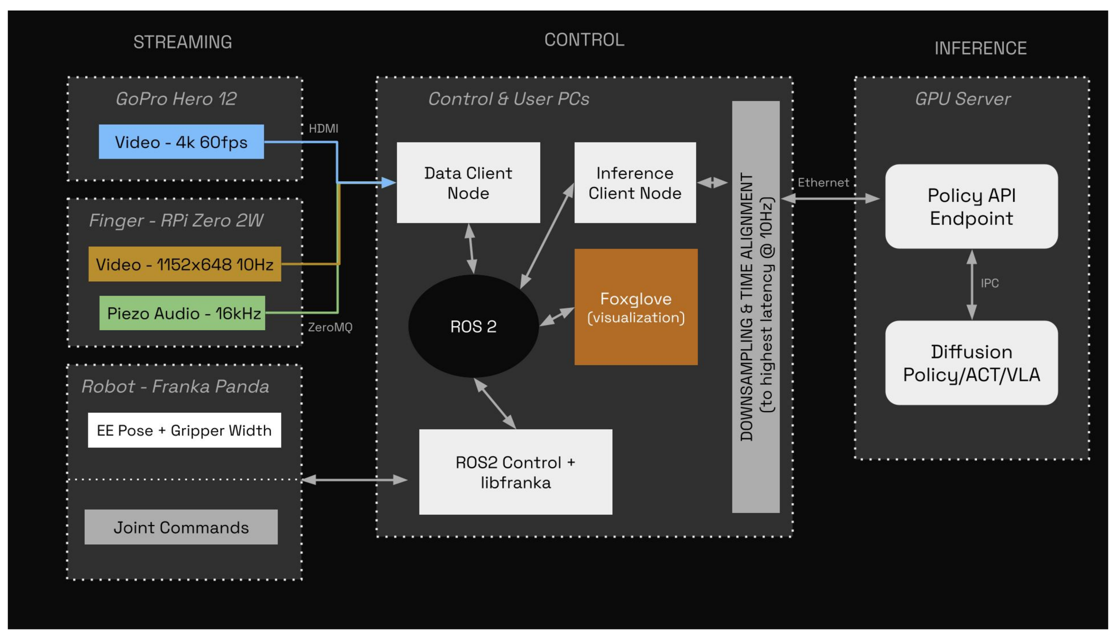
5 - Trajectory Filtering

On a per-embodiment basis, episodes whose trajectories violate the kinematic constraints of the arm are discarded. We filter heuristically via workspace size as in prior works, but evaluation of trajectory feasibility through inverse kinematics is also being explored, enabling experimentation with policies acting in joint space rather than EE pose.



Above: Interactive data visualization using Foxglove is available for both EE livestream and gripper playback.

Policy Training & Deployment



Model Architecture *(in progress)*:

We follow PolyTouch's model architecture [6] as-is for its demonstrated performance on similar inputs. Briefly:
Proprioceptive input is encoded through an MLP, while all other inputs are encoded using pre-trained ViTs (T3 for tactile, AST for audio); these are pooled, combined, and projected into latent observations for a diffusion policy head, which outputs SE(3) EE poses + gripper width.

Training Adjustments *(in progress)*:

We modify PolyTouch's training approach to support a gripper as follows:
1. We follow ManiWAV's approach [3] for addressing the domain gap between gripper and EE piezo audio signals, which is to augment the training data with a background noise dataset and noise sampled from the real robot. **2.** We follow UMI's accepted approach of specifying EE poses for the model relative to the first pose in each action trajectory, rather than in the arm's base frame.

Discussion + Next Steps

Key questions for evaluation: Which sensor modalities matter when, and how do they interact? Do these sensors capture the full "intuitive" value of each modality? (i.e. does the contact mic provide A. nothing, B. a binary signal of contact/no contact, or C. can the model also learn to detect state changes or different materials in manipulanda?) How does the model prioritize conflicting signals (i.e. picking up transparent objects)? To help answer these questions, a set of ablation studies is planned, as well as a set of evaluation tasks designed to illuminate the system's strengths, weaknesses, and boundaries of capability.

Performance objectives in evaluation: Match similar systems (see table) in all evals, while adding performance in dynamic, contact-rich tasks that cannot be demonstrated with teleoperation or performed without touch and fine-grained timing of contact events.

- Contact-rich, dextrous tasks unsuitable for optical-only manipulation (pipetting, glassware handling, screwing a lid on a jar)
- Dynamic tasks infeasible for teleoperation (tossing blocks into a bin)
- Adversarial conditions in which some modalities might fail (ie lighting changes during eval)

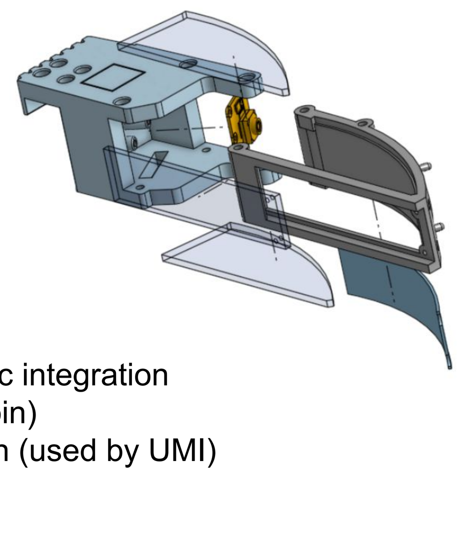
Hardware-software co-design opportunities: One motivation of this work is the belief that as learned models take over increasing portions of the robot control stack, hardware becomes a more fluid design variable. PolyUMI is explicitly designed to support rapid iteration at the hardware level, so that once the learning system is end-to-end integrated and a baseline performance is achieved, incremental hardware development ideas can be developed with rigorous task-based evaluation at every step. Ideas to pursue:

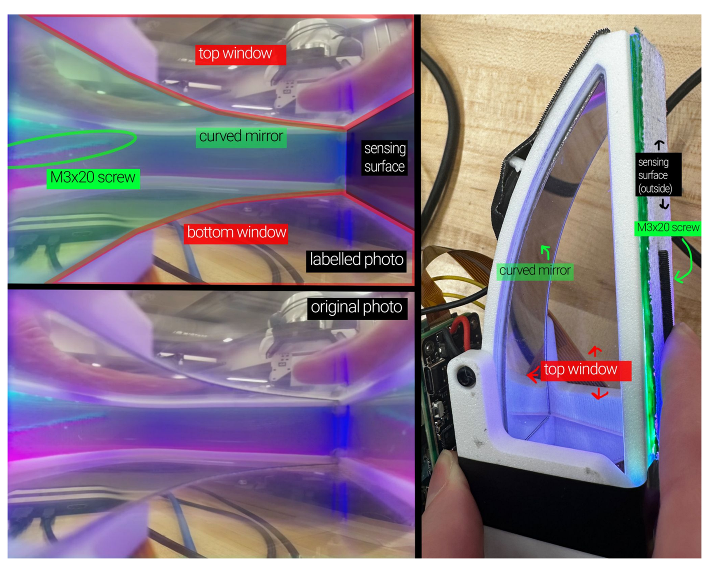
- Optimize mirror geometry - minimize distortion? other heuristics?
- Increase sensitivity & resolution at the fingertip via changes to lighting, mirror, digital preprocessing, etc
- Experiment with mic placement and sensing surface materials

Tactile Sensing

PolyUMI adapts the PolyTouch optical-acoustic finger [6] for a mobile gripper setting. The concept of operation is as follows:

A fisheye camera uses a curved mirror to receive an "overhead view" of a conformant gel-like sensing surface (whose color gradient changes when deformed by pressure from an object), as well as peripheral views of the manipulation area.





Benefits vs. other optical touch sensors (i.e. GelSight):

- Peripheral vision + larger sensing surface area + contact mic integration
- Stiff, sharp fingertip helps with dextrous tasks (i.e pick up coin)
- Compatible with the popular fin-ray conformant finger design (used by UMI)

Novel Contributions:

- Ground-up redesign of PolyTouch finger to enable:
  - Reuse of the same finger between handheld gripper and EE, with fast (<5min) transfer from one to the other
  - Power supplied by same 5V LiPo battery as RPi (reduced LED count, changed placement and type)
- 100% open source (BOM, build instructions, and CAD files all online; PolyTouch is closed source)

ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION 2026 VIENNA

References & Acknowledgements

[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] X. Zhu, B. Huang, and Y. Li, "Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[3] 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.

[4] 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.

[5] Y. Xu, L. Wei, P. An, Q. Zhang, and Y.-L. Li, "exumi: Extensible robot teaching system with action-aware task-agnostic tactile representation," in *Conference on Robot Learning (CoRL)*, 2025.

[6] 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.

[7] 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.

[8] C. Campos, R. Elvira, J. J. Gomez, 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.

[9] S. Urban, "Openicc: An open imu and camera calibrator." https://github.com/urbste/OpenImuCameraCalibrator.

[10] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America A*, vol. 4, no. 4, pp. 629–642, 1987.

[11] J. Zhao, Y. Ma, L. Wang, and E. H. Adelson, "Transferable tactile transformers for representation learning across diverse sensors and tasks," 2024.

[12] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," 2021.

A special thanks to Dr. Matthew Elwin for his support throughout this project, and to Dr. J. Edward Colgate for supervising the next phase of PolyUMI's development and use as a research platform. Thanks as well to Toby Buckley, and to Dr. Yutao He at UCLA, for their advice & feedback.
Finally, thanks to the many incredible contributors in this field who made this work possible.

GitHub Repo + Design Resources

(build your own!)