PolyUMI: Accessible Visual-Tactile-Audio Data Collection for Object Inference and Manipulation

Conor Wood*,[1], Rickmer Krohn*,[2,3,4], Aravind Ramaswami[1], Anunth Ramaswami[1], Nils Dengler[2,3,4], Kevin M. Lynch[1], J. Edward Colgate[1], Georgia Chalvatzaki[2,3,4] and Matthew L. Elwin[1]

***Abstract*— Humans typically rely on vision, touch, hearing, and proprioception to perceive contact and adapt their actions during manipulation. Providing robots with comparable responsiveness therefore requires hardware that can retain and use these complementary sensory signals. Most imitation-learning systems, however, observe demonstrations primarily through vision and proprioception, limiting access to contact information that is difficult to infer visually. We present PolyUMI, an open-source platform for scalable visual–tactile–audio demonstration collection and robot deployment. Its lightweight, wireless handheld gripper records synchronized wrist-camera, optical tactile, contact-audio, and proprioceptive observations without requiring a tethered workstation. The same sensing finger can be transferred to the robot end effector, preserving the sensing geometry between demonstration collection and policy execution. To effectively use these heterogeneous observations, we further introduce VisTA, a token-level multimodal policy that integrates information across sensors and time to predict contact-aware robot actions. Experiments spanning object inference, slip control, and contact-rich manipulation show that touch and audio reveal task-relevant information beyond vision and that VisTA is competitive with or outperforms existing multimodal policies. Together, PolyUMI and VisTA provide an accessible pipeline for collecting multimodal demonstrations and learning policies that perceive physical interaction beyond vision.**

I. Introduction

Learning contact-rich manipulation from demonstration requires policies that can perceive and respond to physical interaction. Here, vision already provides essential information about object pose and scene geometry, but falls short if contact events are occluded or difficult to resolve visually. Humans can instead combine vision with touch and sound to detect slip, recognize surface texture, and determine whether a component has seated correctly. Consequently, manipulation policies trained only on visual and proprioceptive observations lack direct access to these complementary cues, limiting the ability to detect and respond, e.g., to subtle changes in contact.

Prior work has begun to address this sensory gap by incorporating touch and sound into robot learning. Optical tactile sensors capture local surface deformation and contact geometry [1], [2], while contact microphones record vibrations generated during physical interaction [3]. Together with vision, these modalities have supported policies for pouring [4], wrench insertion [5], and dense packing [6]. Making such capabilities broadly accessible requires practical interfaces that capture these complementary signals during human demonstrations. The Universal Manipulation Interface (UMI) [7] provides a foundation for this approach by enabling handheld demonstration collection without robot teleoperation. Subsequent extensions incorporate tactile sensing [8], [9] and contact audio [3], bringing contact information into this workflow. The remaining challenge is to integrate synchronized vision, touch, and audio into a single accessible device while maintaining consistent sensing between demonstration collection and robot deployment. This motivates a unified interface that makes multimodal demonstrations easy to collect and their sensory information directly available during policy execution.

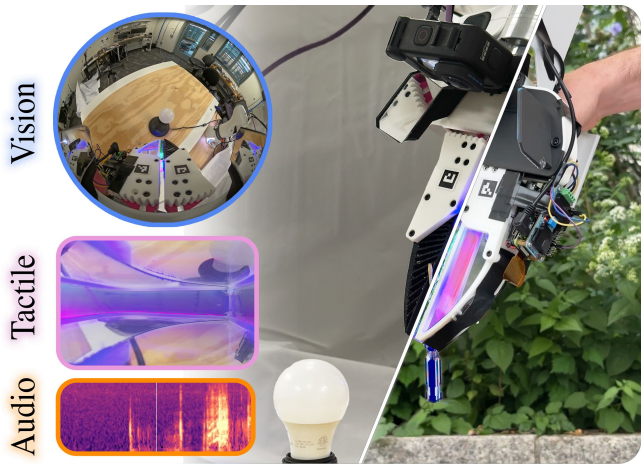

Fig. 1: PolyUMI connects wireless demonstration collection (right) to robot policy execution (left) using a shared multimodal sensing finger. Wrist vision, optical tactile images, and contact audio provide complementary observations for learning contact-rich manipulation.

We address this with PolyUMI, an open-source, wireless interface for synchronized visual–tactile–audio demonstration collection and robot deployment. PolyUMI combines wrist vision and proprioception with an optical tactile finger and integrated contact microphone inspired by PolyTouch [5]. The same sensing finger can be transferred between a self-contained handheld device and a robot gripper, reducing observation shifts between demonstration collection and policy execution. To translate these observations into robot actions, we additionally propose VisTA, a token-level

* Equal Contribution; [1] Center for Robotics and Biosystems, Northwestern Univ., Evanston, IL, USA
[2] Interactive Robot Perception & Learning (PEARL) Lab, TU Darmstadt, Germany; [3] Hessian.AI; [4] Robotics Institute Germany (RIG)
Corresponding author: rickmer.krohn@tu-darmstadt.de

multimodal policy for contact-rich manipulation. Modality-specific input stems encode the sensor observations, while a joint Transformer encoder integrates information across sensors and observation times. The fused representation conditions a flow-matching action model that predicts relative end-effector commands.

We evaluate the contribution of multimodal sensing and the effectiveness of VisTA through object-property inference, closed-loop slip control, and contact-rich manipulation. Sensor ablations show that tactile and auditory observations provide information about visually occluded object properties and support reactive slip control. In manipulation, VisTA outperforms the evaluated baselines on board wiping and matches the strongest multimodal baseline on lightbulb turning. These results highlight both the task-dependent benefits of contact sensing and the importance of combining contact feedback with visual guidance.

In summary, our contributions are:

1) **PolyUMI:** an open-source, wireless platform for synchronized visual–tactile–audio demonstration collection, with a reusable sensing finger shared between handheld and robot-mounted embodiments.
2) **VisTA:** a token-level multimodal policy that combines modality-specific observations through joint transformer fusion and a flow-matching action head.

We release the hardware designs, electronics, firmware, fabrication instructions, and learning software on GitHub.

II. Related Work

A. Multimodal Demonstration Collection

Robot learning requires interfaces that make informative demonstrations easy to collect [2], [7], [9]–[13]. UMI [7] addresses this need through a portable handheld gripper and a corresponding policy-transfer framework. Touch in the Wild [9] extends portable collection with tactile sensing and representation pretraining, while ViTaMIn [12] integrates optical tactile sensing into a gripper interface. TacUMI [8] and UMI-FT [14] additionally incorporate force-related measurements. PolyUMI builds on these efforts by combining synchronized vision, touch, and contact audio in a wireless interface that shares the same sensing finger between collection and deployment.

B. Tactile and Auditory Contact Sensing

Touch and sound expose interaction details that can be difficult to observe with scene cameras. ManiWAV [3] introduces an ear-in-hand interface for learning manipulation from in-the-wild audio-visual demonstrations. PolyTouch [5] combines optical tactile sensing, contact audio, and peripheral vision in a robotic finger for diffusion-policy learning. DIGIT 360 [15] similarly integrates multiple fingertip sensing channels. Our finger follows the PolyTouch sensing principle through an independently developed, open-source implementation designed for handheld use, modular maintenance, and transfer between a robot gripper and a handheld data-collection device.

C. Multimodal Policy Learning

Exploiting complementary sensor observations requires representations that preserve their distinct information while supporting cross-modal reasoning. Visual–tactile learning has explored joint representations [16] and masked multimodal learning [17]. MulSA [6] combines vision, touch, and audio through self-attention, while stage-guided fusion [4] adapts modality weighting to task progress. Sparsh-X [18] learns multisensory touch representations from tactile images, audio, motion, and pressure. VisTA builds on token-level fusion by jointly processing spatial and temporal observations to condition a flow-matching action head. Comparisons with alternative fusion methods and sensor ablations examine the contributions of sensing and policy architecture.

III. Visual-Tactile-Audio Data Collection for Object Inference and Manipulation

Contact-rich manipulation depends on information that is often occluded from vision and revealed only through deformation, vibration, and sound. Exploiting these signals requires reliable synchronized data collection across multiple sensors and effective multimodal fusion. We address these challenges with PolyUMI, a wireless parallel gripper that reuses the same sensing finger across handheld demonstration collection and robot execution, providing an accessible and scalable option for multimodal data acquisition and policy deployment. Furthermore, to effectively use PolyUMI, we propose VisTA, a token-level fusion policy that jointly reasons over visual, tactile, auditory, and proprioceptive observations to generate contact-aware actions for complex manipulation tasks. In this section, we first describe the PolyUMI system, its modular sensing finger, and the workflow used to collect and deploy demonstrations. We then detail the synchronization and preprocessing of the individual sensor streams. Finally, we introduce VisTA's observation and action representations, token-level fusion architecture, and flow-matching policy head.

A. The PolyUMI System

*1) **System Overview**:* PolyUMI comprises two complementary embodiments: (1) a handheld gripper for collecting demonstrations and (2) a robot-mounted end-effector for policy execution. The handheld device follows the UMI paradigm [7], extended with tactile and auditory contact sensing. Both embodiments use the same transferable multimodal finger, thereby reducing observation shifts between data collection and deployment. The handheld device includes an onboard battery, computer, and audio interface, eliminating the need for a tethered workstation and enabling practical multimodal data collection in unstructured environments. The data collection software launches automatically and is controlled through a button and status LED, while a wrist mounted camera is triggered over Bluetooth to coordinate recording across sensor streams. For deployment, we provide a data streaming application and a ready-to-use mounting interface for the Franka Hand in our open-source release, while the modular design allows PolyUMI to be adapted

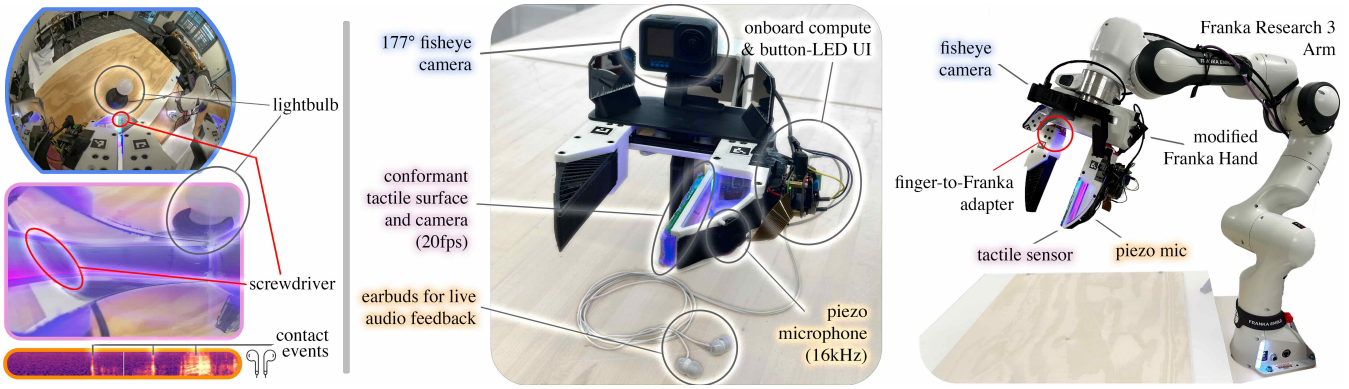

Fig. 2: *Left*: The relationship between different sensor views of the same scene. The tactile camera "sees" the screwdriver through the sensing surface, as well as the lightbulb through peripheral vision. The wrist camera sees the lightbulb directly, while the screwdriver is mostly occluded by the fingers holding it. As the device interacts with its environment, contact events are detected by the piezo microphone, which can be seen as spikes in the audio spectrogram, as well as heard through headphones during gripper operation. *Center, Right*: The PolyUMI system is composed of a gripper and end-effector, carefully designed to share the same tactile finger component and task-oriented geometry.

	Sensor modalities					
Method	Vision	Tactile	Audio	Open source	Cable free	Live feedback
UMI [7]	✓	×	×	✓	✓	×
FastUMI [13]	✓	×	×	✓	×	×
ViTaMin [12]	✓	✓	×	×	✓	×
Actuated UMI [11]	✓	✓	×	✓	✓	×
ManiWAV [3]	✓	×	✓	✓	✓	×
TacThru [2]	✓	✓	×	✓	×	×
Touch in the Wild [9]	✓	✓	×	✓	×	✓
PolyUMI (Ours)	✓	✓	✓	✓	✓	✓

TABLE I: Comparison of multimodal data collection systems. PolyUMI is the only interface able to collect synchronous vision, tactile and audio data. It also supplies a novel modality of operator feedback through live audio monitoring. The system's hardware and software are open-sourced for community use.

to other robotic grippers by replacing only the platform-specific mount. Table I compares PolyUMI with existing multimodal demonstration interfaces and shows that it is the only interface that combines synchronized vision, touch, and contact audio in a handheld platform. An Overview of the proposed hardware is illustrated in Fig. 2

*2) **Modular Multimodal Sensing Finger**:* The multimodal finger integrates an optical tactile sensor and a contact microphone into a compact module. Its sensing principle is inspired by PolyTouch [5]. However, since the original implementation is not publicly available, we independently developed the mechanical structure, electronics, firmware, and fabrication process, and adopted several alternative design choices which can be seen in the open source release, to make the sensor softer and easier to manufacture. We also added dynamic lighting compensation to enable the use of the finger across multiple environments, necessary for an in-the-wild data collector. The output specifications of PolyUMI's modalities are summarized in Table II. In the following we will explain each part of the system in more detail.

*a) **Optical Tactile Sensing**:* An internal camera observes a deformable reflective surface through a curved mirror, providing a near-overhead view of a large contact region. In comparison to PolyTouch (which uses 3 layers),

Modality	Sensor	Output
Tactile	Optical finger	20 fps, 1152×648 MJPEG
Audio	Contact mic	16 kHz mono PCM
Vision	GoPro Hero 12 + fisheye	60 fps, 1920×1080 MP4
Proprioception	SLAM / ArUco	6-DoF pose + gripper width

TABLE II: PolyUMI sensing modalities and their native output specifications.

our sensing surface consists of seven layers of VHB tape coated with aluminum powder and sealed with medical tape. The layered construction increases conformability, while the reflective coating makes local deformation visible to the internal camera. The tactile stream is recorded at 20 fps with a resolution of 1152×648. Camera parameters are adjusted online to maintain usable observations under changes in ambient illumination during in-the-wild data collection.

*b) **Contact Audio**:* A contact microphone is rigidly coupled to the sensing-finger. In contrast to an external microphone, it primarily records vibrations transmitted through the finger and grasped object. These vibrations provide temporally precise cues for events such as contact onset, sliding, impact, and slip [3]. Audio is sampled as a 16 kHz mono PCM signal through a Raspiaudio ULTRA+ interface. A headphone output additionally allows the demonstrator to monitor the contact-audio signal while collecting data, which led to more consistent demonstrations during contact-phases.

*c) **Wrist Vision**:* PolyUMI uses a GoPro Hero 12 with a MAX Lens Mod 2.0, providing an approximately 177° field of view. Side mirrors expose additional views of the gripper and manipulated object within the egocentric image. The camera records 1920×1080 video at 60 fps. During preprocessing, the video is decoded and sampled at timestamps aligned with the policy observations.

*d) **Proprioception**:* The handheld and robot-mounted embodiments obtain proprioception differently but express it in a common end-effector representation. For handheld demonstrations, monocular-inertial ORB-SLAM3 [19] estimates the six-degree-of-freedom gripper trajectory from the GoPro video and IMU (via a custom fork which supports more precise gripper masking during mapping). ArUco

	Robot End-Effector [7]	Handheld UMI Gripper [7]	Multimodal Sensing Finger
Hardware cost	\$742.48	\$700.27	\$235.96
Assembly time	30 min	2 h	4 h

TABLE III: Estimated hardware cost and manual assembly time for the main PolyUMI components. The multimodal sensing finger adds approximately \$240 and four hours of assembly relative to the original UMI platform.

markers attached to the fingers provide the gripper width. During robot execution, the end-effector pose is computed from the robot joint state, while the gripper width is read directly from the hand.

*3) **Manufacturing and Assembly**:* PolyUMI is designed for straightforward manufacturing, assembly, and maintenance. As summarized in Table III, adding the multimodal sensing finger to the original UMI platform requires $\approx$ \$240 in additional hardware and four hours of assembly. The complete design uses modular components and minimal software dependencies, reducing the setup effort required to deploy the system. Importantly, the finger can be easily transferred between the gripper and the robot so that policies can be trained and deployed on a single sensor, minimizing domain shift. Transferring the finger between the two embodiments takes ≈ 10 minutes.

*4) **Synchronization and Data Processing**:* PolyUMI's sensors operate at different sampling rates and latencies. We associate all observations with a common trajectory timeline and construct policy inputs at a 10 Hz control frequency. All observations are interpolated to the timestamp of the newest observation from the stream with the highest latency. Additionally, the first few actions in each chunk are skipped in execution, to compensate for the total combined latency from the synchronized observations to to the action execution. The exact number of actions skipped is calculated through a mix of online measurement and offline calibration. At each policy step, the policy receives the current and previous wrist-camera and tactile images, resulting in an observation history of $H = 2$. Both image modalities are resized to 224×224 pixels before being provided to the model. The continuous 16 kHz audio signal is divided into windows aligned with the policy observations. Each window is converted into a log-Mel spectrogram using 128 Mel bands, a 25 ms analysis window, and a 10 ms hop. We retain 48 spectrogram frames, corresponding to approximately 0.5 s of recent contact audio. The resulting audio observation has shape $3 \times 128 \times 48$. This processing produces aligned visual, tactile, auditory, and proprioceptive observations while retaining the temporal resolution required to capture short contact events.

*5) **Control and Execution**:* Following UMI [7], each forward pass of the policy outputs a sequence of $\mathrm{SE}(3)$ positions in gripper frame. On the arm, these 10 Hz commands are interpolated and translated into joint torques at 1 kHz by a cartesian impedance controller (adapted from [20]) to achieve compliant manipulation. To achieve precise gripper control, the commercially available Franka Hand was modified to support 200 Hz continuous position control, whereas the off-the-shelf version does not support real-time use.

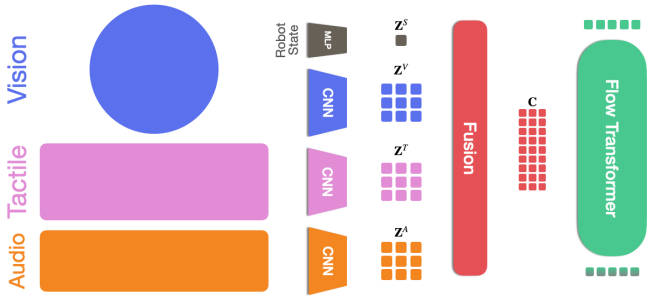

Fig. 3: The VisTA model consists of sensor-specific CNN encoders and joint self-attention fusion. The high number of fused conditioning tokens $\mathbf{C}$, then conditions a flow matching transformer.

B. VisTA Multimodal Policy

Observations obtained by PolyUMI provide complementary information at different spatial and temporal scales. Wrist vision captures the global scene, tactile images describe local contact geometry, contact audio captures short interaction events, and proprioception represents the robot state. Effectively combining these signals requires preserving their modality-specific structure while allowing information to be exchanged across sensors and time. We therefore propose **VisTA**, a token-level multimodal fusion architecture which enables the policy to dynamically combine global **Vis**ual context with local **T**actile and **A**uditory feedback in order to solve contact-rich manipulation tasks. Given a small history of multisensory obervations $\mathcal{O}_t = \left\{\mathbf{o}^V_{t-H+1:t}, \mathbf{o}^T_{t-H+1:t}, \mathbf{o}^A_t, \mathbf{s}_{t-H+1:t}\right\}$, and the robot state $\mathbf{s}_t = \left[\mathbf{p}_t, \mathbf{r}_t, g_t, \mathbf{r}^{\mathrm{rel}}_t\right] \in \mathbb{R}^{16}$, the policy predicts an action chunk of $\mathbf{x} \in \mathbb{R}^{T_a \times d_a}$ with $T_a = 16$ and $d_a = 10$. Each action contains a 3D translation, a 6D rotation, and a relative gripper-width command. The pose commands are expressed relative to the current end-effector pose, making the action representation independent of the global workspace frame. An Overview of the VisTA encoding and policy head is given in Figure 3.

Each modality of the observation O_t is processed to form embeddings of $D = 624$. Wrist camera and tactile images are encoded by a CNN with a stride of 32, resulting in 98 tokens over $H = 2$ timesteps. Modality-specific spatial embeddings and shared temporal embeddings retain the origin and time index of each token. The log-Mel spectrogram is encoded by an audio CNN to 96 tokens, while the robot state is independently encoded via an MLP. The full observation $\mathbf{Z}_0 = \left[\mathbf{Z}^V; \mathbf{Z}^T; \mathbf{Z}^A; \mathbf{Z}^S\right] \in \mathbb{R}^{N \times D}$ leads to $N = 294$ tokens. An eight-layer pre-normalized transformer with eight attention heads (f_{fusion}) jointly refines this sequence by bi-directional self-attention and MLP subblocks. The attention mechanism operates over the complete token sequence, leading to direct information exchange across modalities, spatial locations, and observation times. In contrast to fusing one pooled feature per sensor, this design preserves local tactile deformation, short auditory events, and visual scene to condition the policy head. The fused representation $\mathbf{C} = f_{\mathrm{fusion}}(\mathbf{Z}_0)$ conditions a DiT-style [21] action model with 18 transformer layers via cross attention. The model

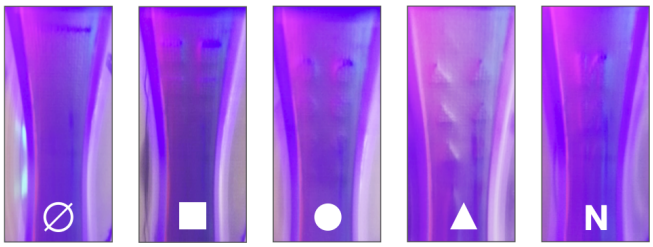

Fig. 4: Tactile images of the shape recognition objects. From left to right: *Flat*, *Square*, *Circle*, *Triangle*, and the letter *N*.

predicts a conditional velocity field $\hat{\mathbf{v}}_\theta(\mathbf{x}_\tau, \tau, \mathbf{C})$ over the complete action chunk and is trained using conditional flow matching [22], [23]. The continuous flow time τ is injected into each DiT block using AdaLN-Zero conditioning [24].

IV. Experiments

We evaluate PolyUMI along three claims that motivate our system. First, we ask whether its tactile and contact-audio streams can be used to expose object properties that are difficult to infer from wrist vision alone. Second, we test whether these modalities provide precise feedback for closed-loop slip control. Third, we evaluate whether the resulting multimodal observations improve imitation policies on contact-rich manipulation tasks. We use V, T, and A to denote wrist vision, optical tactile images, and contact audio, respectively.

A. Experimental Procedure

For all experiments, we use a Franka FR3 equipped with our PolyUMI end-effector. Demonstrations and policy rollouts therefore share the same sensing geometry. Policies predict an action chunk containing 16 future actions. The first three actions are executed with temporal ensemble [25] before the policy receives a new observation and replans. Within each experiment, all baselines and modality ablations use the same demonstrations, train–test split, observation and action horizons, input preprocessing, data augmentation, and optimization budget. We train every model for 120 epochs and select the final checkpoint for evaluation. For manipulation and control tasks, we report the number of successful rollouts over the total number of evaluation trials. For classification tasks, we report accuracy on the held-out test set.

B. Object Properties and Slip Control

We first evaluate whether PolyUMIs modalities capture task-relevant information arising during contact, before assessing their impact on longer-horizon manipulation.

1) ***Tactile Shape Recognition****:* Fine surface geometry is difficult to recover visually during manipulation, as the contact region is occluded by the gripper and the relevant features are smaller than the effective resolution of an external camera. We therefore investigate whether PolyUMI's tactile sensor captures geometry-specific deformation signatures that are sufficiently informative to distinguish surface features at the contact interface. We formulate this as a classification task over five 3D-printed patterns, i.e., *Flat*, *Square*, *Circle*, *Triangle*, and the letter *N*, and train a classifier to predict the pattern from a single tactile reading. The tactile readings are shown in Fig. 4 while the results are displayed in Fig. 5.

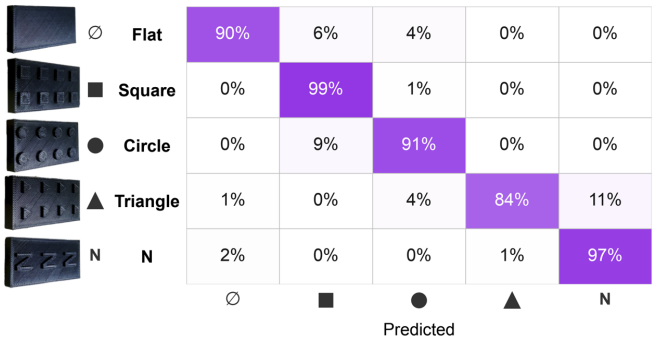

Fig. 5: Tactile shape classification results. PolyUMIs tactile sensor enables object recognition across 5 classes.

We collect 27 scenes per class under varying backgrounds, lighting conditions, and in-hand object poses. Each scene contributes approximately 20 samples, resulting in a total of 135 unique settings and 2700 tactile images. We split the data at the scene level: the classifier is trained on 1986 samples from 20 scenes per class and evaluated on the 714 samples from the 7 held-out scenes per class, ensuring that no scene appears in both splits. The model consists of a CNN-based tactile encoder followed by an MLP head that predicts the class label from a single tactile image. Training runs for 120 epochs and takes approximately 11 minutes on a GeForce RTX 4060. Fig. 5 reports the resulting confusion matrix. The classifier reaches an accuracy of 92.3 %, indicating that PolyUMI's optical tactile sensor provides features that are sufficiently discriminative for tactile shape recognition.

2) ***Object-in-Box Classification****:* Many object properties remain initially hidden when objects need to be manipulated that are not fully hollow or filled, but contain objects inside. Humans can handle these scenarios and infer a container's contents by shaking it and listening to the resulting impacts. These sounds depend on properties such as the objects' mass, geometry, and material, whereas vision can observe only the motion of the container and robot. We therefore investigate whether PolyUMI's contact-audio sensor captures object-specific collision signatures that are sufficiently informative to classify visually occluded objects.

Inspired by [26], we place one of three object classes, i.e., gears, thin screws, or thick screws, inside the same closed box. At the beginning of each demonstration, the box position is randomized over a 6 cm range along the x-, y-, and z-axes. As shown in Fig. 6, the robot then shakes the box for 20 oscillations by following a sinusoidal trajectory with an amplitude of 10 cm and a period of 0.15 s. This controlled motion allows us to compare the sensory signals generated by the different object classes while varying the initial configuration of the objects inside the box.

We collect 30 demonstrations for each object class, using 20 demonstrations for training and reserving 10 for validation. Each available modality is processed by a separate CNN encoder. The resulting modality features are concatenated

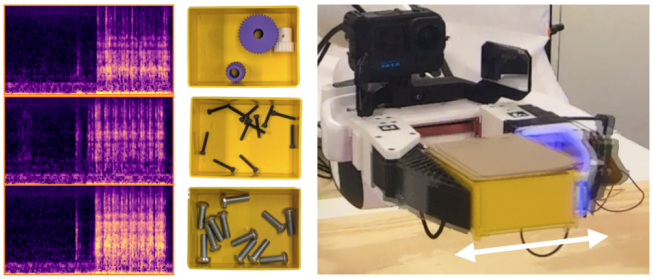

Fig. 6: Audio signals (left) generated by shaking gears, thin screws, and thick screws (middle) with the robot arm (right) enable classification of visually occluded objects, substantially outperforming vision and touch.

and passed to an MLP classification head that predicts one of the three object classes. To account for variability during training, we repeat each experiment with three random seeds and select the checkpoint with the highest validation accuracy. The confusion matrices therefore aggregate 30 predictions per object class and 90 predictions in total for each sensor configuration.

	Acc.	Gears			Thin Screws			Thick Screws		
V	44%	43%	47%	10%	53%	20%	27%	10%	20%	70%
T	44%	37%	37%	27%	33%	20%	47%	17%	13%	70%
V+T	42%	17%	70%	13%	27%	53%	20%	7%	37%	57%
A	80%	77%	23%	0%	17%	77%	7%	0%	13%	87%
V+A	81%	63%	37%	0%	10%	83%	7%	0%	3%	97%
T+A	82%	73%	27%	0%	23%	73%	3%	0%	0%	100%
V+T+A	79%	77%	23%	0%	37%	60%	3%	0%	0%	100%
	Acc.	Gears	Thin Screws	Thick Screws	Gears	Thin Screws	Thick Screws	Gears	Thin Screws	Thick Screws

Fig. 7: Prediction distributions for object-in-box classification, aggregated over three random seeds. Each panel corresponds to one ground-truth object class, and each row represents a sensor configuration. The green outlined cells indicate correct predictions.

Figure 7 shows that models without audio, i.e., vision-only (44 %), tactile-only (44 %) and visual–tactile (42 %), achieve only 10 % higher accuracy than random guessing (33 %). Notably, the vision-only model identifies thick screws more reliably than the other classes, reaching 70 % accuracy. This suggests that the model can partially exploit differences in the visible arm and container motion caused by the greater mass of the thick screws. However, visual observations alone do not provide sufficient information to distinguish the three classes reliably. Consequently, adding contact audio substantially improves classification leading to $\approx$ 80.0 % in every sensor configuration. In particular, the two tactile-audio conditioned models classify thick screws with 100 % accuracy, while The visual–audio model correctly identifies thin screws in 83 % of the trials. These results demonstrate that the structured audio recorded by PolyUMI contains object-specific information that is not available from vision or tactile images alone and enables the classification of objects that remain occluded inside the box with only 20 training demonstrations per class.

*3) **Closed-Loop Slip Control**:* Finally, successful manipulation requires the robot to also respond to contact events as they evolve. In this context, object slippage is particularly challenging to control because it can be subtle, visually occluded, or too rapid to detect reliably from wrist-camera

Slip Control	
Sensors	SR
V+T+A	**8/10**
T+A	7/10
V+T	3/10
V+A	3/10
A	0/10
T	3/10
V	2/10

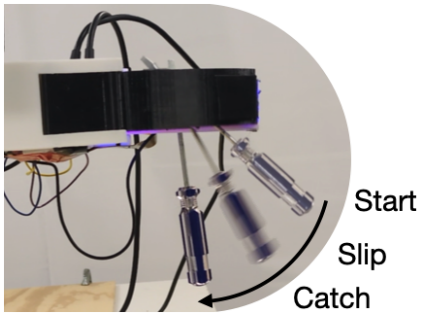

Fig. 8: Sensor ablation success rates (SR) across 10 trials (left) and side view of the Slip Control Task setup (right). The combination of tactile and audio sensing enables closed-loop slippage control, where other sensor modalities fail.

observations alone. Tactile deformation and structure-borne vibrations, in contrast, provide direct measurements of relative motion between the object and gripper. We therefore investigate whether PolyUMI's tactile and contact-audio modalities provide sufficiently precise feedback for closed-loop slip control.

In this experiment, the robot initially holds a screwdriver approximately horizontal, at 90° from the downward direction (see Fig. 8). The policy's goal is to rotate the screwdriver to 0° (approximately vertical) by modulating the grasping force, without allowing the screwdriver to fall. This task requires a careful balance, as tightening too early prevents the screwdriver from reaching the target orientation, whereas reacting too late causes the object to slip out of the gripper. Performing this task via teleoperation without haptic feedback is particularly hard due to the fine motor control required, making our handheld PolyUMI platform essential to train a successful policy. In this experiment, a rollout is considered successful if the screwdriver remains in the gripper with a final orientation between 0° and 10° from the goal position.

For this task, we collect 80 demonstrations and encode the visual and tactile frames using ResNet backbones, while contact audio is processed using a temporal CNN. The resulting features condition a diffusion-policy U-Net that predicts relative changes in gripper width. By comparing the full visual–tactile–audio policy with different modality ablations, we assess whether contact sensing enables more accurate and timely regulation of object slip. The full V+T+A policy succeeds in 8 out of 10 trials, compared to 2 out of 10 trials for the vision-only policy, as shown in Fig. 8. These results suggest that contact-based observations provide important feedback for controlling fine-grained gripper–object interactions.

C. Contact-Rich Manipulation

The preceding experiments isolate each modality of our PolyUMI platform and highlight that it can perceive object properties and detect dynamic contact events. However, the central question is whether these multimodal observations also improve complete manipulation behaviors. In contact-rich tasks, a policy must transition between free-space motion and physical interaction, maintain appropriate contact, and

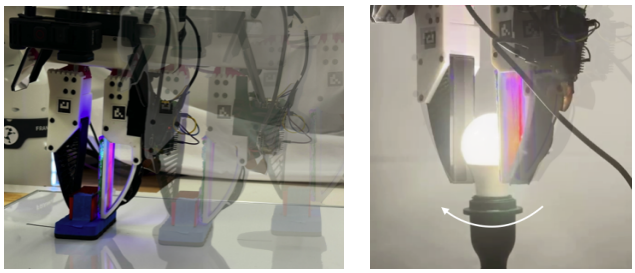

Fig. 9: Our contact-rich manipulation tasks Board Wiping (left) and Lightbulb Turning (right).

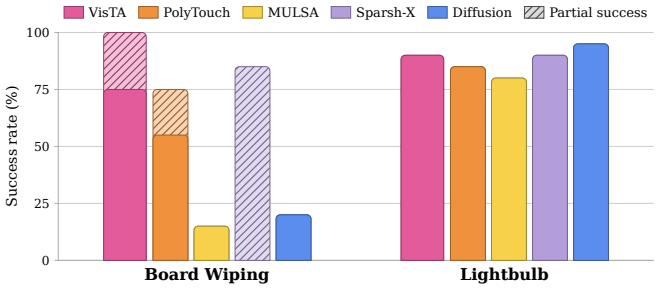

Fig. 10: Manipulation results of Board Wiping and Lightbulb. VisTA outperforms state of the art visual tactile audio models in the board wiping task. All models are able to solve Lightbulb turning. while multisensor models fail to detect task stages, rather than contact modes.

respond to small changes that may be difficult to observe visually. Errors during these phases can accumulate over time, causing otherwise successful trajectories to fail during alignment, insertion, or surface following. We therefore evaluate whether the tactile and auditory information captured by PolyUMI improves end-to-end manipulation reliability.

Here, we consider two tasks that represent distinct forms of physical interaction. First, a **board wiping** task requires sustained contact and continuous surface following. Second, a lightbulb connection task combines the three difficult behaviors of alignment, insertion, and rotational engagement. All policies receive proprioceptive information together with the sensor observations. An overview of the task setups can be seen in Figure 9. Together, these tasks allow us to evaluate whether multimodal policies can use contact information across different interaction types rather than only in a single controlled setting.

*1) **Baselines and Implementation**:* To determine whether performance gains arise from the additional contact modalities or from the way in which they are fused, we compare VisTA against both a vision-only policy and three multimodal fusion baselines. The vision-only **Diffusion Policy** [7], [27] uses a ResNet visual encoder and a U-Net policy head, providing a reference for manipulation without tactile or auditory observations. The remaining baselines receive the same visual, tactile, auditory, and proprioceptive inputs as VisTA but employ different strategies for encoding and combining them. **MulSA** [6] processes each modality with a separate ResNet-18 encoder [28] and applies attention across modalities and time. We replace the original MLP prediction head with a U-Net diffusion model to match other baselines. **Sparsh-X** [18] represents the sensor observations as tokens using modality-specific patch stems and exchanges information through attention bottlenecks [29]. The fused tokens are summarized through attention pooling [30] and passed to a DiT policy head trained using flow matching [23]. **PolyTouch** [5] instead relies on pretrained modality-specific representations, using CLIP for vision [31], T3 for tactile sensing [32], and an Audio Spectrogram Transformer for contact audio [33]. Its features are combined through cross-attention and concatenated with the audio class token before conditioning the diffusion policy head. This controlled comparison allows us to separately assess the value of contact sensing and the effect of the multimodal fusion architecture.

*2) **Board Wiping**:* Board wiping tests whether a policy can establish and maintain sustained contact while following a visually defined trajectory. Although the line to be erased is visible, successful execution cannot be achieved through visual tracking alone, as the policy must also regulate the interaction between the eraser and the board throughout the motion. Insufficient contact leaves parts of the line unerased, whereas deviations from the line produce only partial task completion. This task therefore evaluates whether multimodal contact observations help the policy coordinate contact maintenance with visually guided surface following. In each trial, the robot must erase a line of ≈ 25 cm in length. The position of the line is slightly randomized between trials, while the robot begins from the same initial configuration. We consider a rollout fully successful if the complete line is erased and partially successful if the robot maintains contact with the board but fails to erase the entire line. Losing board contact during the task is considered a failure.

The results are shown in Fig. 10. The vision-only Diffusion Policy and MulSA frequently lose contact with the board during the wiping motion, leaving substantial portions of the line unerased. Sparsh-X generally maintains contact but often deviates from the line, resulting primarily in partial successes. In contrast, VisTA reliably establishes contact, maintains it throughout the wiping motion, and follows the line closely enough to erase it completely. These results indicate that successful board wiping requires both contact regulation and spatial tracking, and that VisTA effectively integrates the corresponding tactile, auditory, and visual information.

*3) **Lightbulb Turning**:* Lightbulb turning evaluates whether a policy can coordinate grasping, axial contact, and rotational motion across a multi-stage manipulation sequence. In each trial, the robot begins with a loosely inserted lightbulb and must rotate it until it is fully seated and switches on. We randomize the required bulb rotation and vary the initial vertical gripper position by ± 1 cm. The policy must maintain a stable grasp, reposition the gripper when necessary, and regulate the bulb's axial position while rotating it. A rollout is considered successful if the bulb switches on within 90 s and remains seated after the gripper releases it.

As shown in Fig. 10, all methods achieve success rates of at least 80%. The vision-only Diffusion Policy performs best, closely followed by VisTA and Sparsh-X. The remaining

failures occur primarily during transitions between rotation stages. MulSA occasionally loses the bulb while re-grasping, whereas PolyTouch sometimes retreats before the bulb is fully seated. These results indicate that the task is largely solvable from visual observations because the bulb state and task completion provide clear visual cues. Consequently, tactile and auditory sensing offer limited additional benefit in this setting. Nevertheless, VisTA matches the strongest multimodal baseline and remains close to the vision-only policy, showing that its multimodal fusion supports multi-stage manipulation without compromising performance when vision provides the dominant task information.

V. Conclusion

In this work, we introduced PolyUMI, an open-source, wireless platform for synchronized visual–tactile–audio demonstration collection and robot deployment. By sharing the same sensing finger between handheld and robot-mounted embodiments, PolyUMI preserves the sensing configuration while enabling data collection without a tethered workstation. We further proposed VisTA, a token-level multimodal policy that combines information across sensors and time to generate contact-aware actions. Experiments on object-property inference demonstrated that tactile observations can be used to capture surface geometry, while contact audio reveals hidden object properties. Sensor ablations further showed that combining tactile and auditory feedback improves closed-loop slip control over vision alone. Finally, real-robot manipulation experiments demonstrated that VisTA outperforms the evaluated baselines on board wiping by coordinating sustained contact with visual tracking, while matching the strongest multimodal baseline on lightbulb turning. Together, these results highlight the complementary value of contact sensing and effective multimodal fusion for learning manipulation skills beyond purely visual feedback.

VI. Acknowledgments

We thank Han Zhengxiao for his invaluable advice at the outset of this project, and Toby Buckley and Yutao He for insightful conversations throughout. This research is funded by the German Research Foundation (DFG) Emmy Noether Programme (CH2676/1-1), the EU's Horizon Europe project ARISE (Grant no.: 101135959), the German Federal Ministry of Education and Research (BMBF) project "RiG" (Grant no.: 16ME1001) and the European Research Council (ERC) project "SIREN" (Grant No.: 101163933). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universit¨at Erlangen-N¨urnberg (FAU).

References

[1] 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.

[2] Y. Li, Y. Chen, Z. Zhao, P. Li *et al.*, "Simultaneous Tactile-Visual Perception for Learning Multimodal Robot Manipulation," *IEEE Robotics and Automation Letters*, vol. 11, no. 4, pp. 5254–5261, 2026.

[3] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy *et al.*, "ManiWAV: Learning Robot Manipulation from In-the-Wild Audio-Visual Data," in *Proc. of The 9th Conference on Robot Learning*, 2025, pp. 947–962.

[4] R. Feng, D. Hu, W. Ma, and X. Li, "Play to the Score: Stage-Guided Dynamic Multi-Sensory Fusion for Robotic Manipulation," in *Proceedings of The 8th Conference on Robot Learning*, 2024.

[5] J. Zhao, N. Kuppuswamy, S. Feng, B. Burchfiel *et al.*, "Polytouch: A Robust Multi-Modal Tactile Sensor for Contact-Rich Manipulation Using Tactile-Diffusion Policies," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2025.

[6] H. Li, Y. Zhang, J. Zhu, S. Wang *et al.*, "See, Hear, and Feel: Smart Sensory Fusion for Robotic Manipulation," in *Proc. of The 6th Conference on Robot Learning*, 2022.

[7] C. Chi, Z. Xu, C. Pan, E. Cousineau *et al.*, "Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots," in *Proc. of Robotics: Science and Systems (RSS)*, 2024.

[8] T. Cheng, K. Chen, L. Chen, L. Zhang *et al.*, "TacUMI: A Multi-Modal Universal Manipulation Interface for Contact-Rich Tasks," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2026.

[9] X. Zhu, B. Huang, and Y. Li, "Touch in the Wild: Learning Fine-Grained Manipulation with a Portable Visuo-Tactile Gripper," *Advances in Neural Information Processing Systems*, vol. 38, 2026.

[10] Y. Xu, L. Wei, P. An, Q. Zhang *et al.*, "exUMI: Extensible Robot Teaching System with Action-aware Task-agnostic Tactile Representation," in *Proc. of The 9th Conference on Robot Learning*, 2025, pp. 2536–2554.

[11] E. Helmut, N. Funk, T. Schneider, C. de Farias *et al.*, "Tactile-Conditioned Diffusion Policy for Force-Aware Robotic Manipulation," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2026.

[12] F. Liu, C. Li, Y. Qin, J. Xu *et al.*, "Vitamin: Learning Contact-Rich Tasks Through Robot-Free Visuo-Tactile Manipulation Interface," *arXiv preprint 2504.06156*, 2025.

[13] Z. Zhaxizhuoma, K. Liu, C. Guan, Z. Jia *et al.*, "FastUMI: A Scalable and Hardware-Independent Universal Manipulation Interface with Dataset," in *Proc. of The 9th Conference on Robot Learning*, 2024, pp. 3069–3093.

[14] H. Choi, Y. Hou, C. Pan, S. Hong *et al.*, "In-the-Wild Compliant Manipulation with UMI-FT," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2026.

[15] M. Lambeta, T. Wu, A. Sengul, V. R. Most *et al.*, "Digitizing touch with an artificial multimodal fingertip," *arXiv preprint 2411.02479*, 2024.

[16] M. A. Lee, Y. Zhu, P. Zachares, M. Tan *et al.*, "Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 582–596, 2020.

[17] C. Sferrazza, Y. Seo, H. Liu, Y. Lee *et al.*, "The Power of the Senses: Generalizable Manipulation from Vision and Touch through Masked Multimodal Learning," in *Proc. of the IEEE Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2024.

[18] C. Higuera, A. Sharma, T. Fan, C. K. Bodduluri *et al.*, "Tactile Beyond Pixels: Multisensory Touch Representations for Robot Manipulation," in *Proc. of The 9th Conference on Robot Learning*, 2025.

[19] C. Campos, R. Elvira, J. J. Gómez, J. M. M. Montiel *et al.*, "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.

[20] J. Luo, Z. Hu, C. Xu, Y. L. Tan *et al.*, "SERL: A Software Suite for Sample-Efficient Robotic Reinforcement Learning," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2024.

[21] W. Peebles and S. Xie, "Scalable Diffusion Models with Transformers," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023.

[22] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel *et al.*, "Flow Matching for Generative Modeling," in *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2023.

[23] K. Black, N. Brown, J. Darpinian, K. Dhabalia *et al.*, "$\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization," in *Proc. of The 9th Conference on Robot Learning*, 2025.

[24] S. Dasari, O. Mees, S. Zhao, M. K. Srirama *et al.*, "The Ingredients for Robotic Diffusion Transformers," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2025.

[25] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in *Proc. of Robotics: Science and Systems*, 2023.

[26] Z. Xu, Z. Si, K. Zhang, O. Kroemer *et al.*, "A Multi-Modal Tactile Fingertip Design for Robotic Hands to Enhance Dexterous Manipulation," *arXiv preprint 2510.05382*, 2025.

[27] C. Chi, Z. Xu, S. Feng, E. Cousineau *et al.*, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," *The Intl. Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[29] A. Nagrani, S. Yang, A. Arnab, A. Jansen *et al.*, "Attention Bottlenecks for Multimodal Fusion," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 14 200–14 213.

[30] X.-K. Chen, M. Ding, X. Wang, Y. Xin *et al.*, "Context Autoencoder for Self-supervised Representation Learning," *International Journal of Computer Vision*, vol. 132, pp. 208 – 223, 2022.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. of the IEEE Intl. Conf. on Machine Learning*. PmLR, 2021, pp. 8748–8763.

[32] J. Zhao, Y. Ma, L. Wang, and E. Adelson, "Transferable Tactile Transformers for Representation Learning Across Diverse Sensors and Tasks," in *Proc. of The 8th Conference on Robot Learning*, 2024.

[33] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Interspeech 2021*, 2021, pp. 571–575.