A-SLIP: Acoustic Sensing for Continuous In-hand Slip Estimation

Uksang Yoo[*,1], Yuemin Mao[*,1], Jean Oh[1], Jeffrey Ichnowski[1]

***Abstract*— Reliable in-hand manipulation requires accurate real-time estimation of the slip between a gripper and a grasped object. Existing tactile sensing approaches based on vision, capacitance, or force-torque measurements face fundamental trade-offs in form factor, durability, and their ability to jointly estimate slip direction and magnitude. We present A-SLIP, a multi-channel acoustic sensing system integrated into a parallel-jaw gripper for estimating continuous slip in the grasp plane. The A-SLIP sensor consists of piezoelectric microphones positioned behind a textured silicone contact pad to capture structured contact-induced vibrations. The A-SLIP model processes synchronized multi-channel audio as log-mel spectrograms using a lightweight convolutional network, which jointly predicts the presence, direction, and magnitude of the slip. Across experiments with robot- and externally-induced slip conditions, the finetuned four-microphone configuration achieves a mean absolute directional error of 14.1 degrees, outperforms baselines by up to 12% in detection accuracy, and reduces directional error by 32%. Compared with single-microphone configurations, the multi-channel design reduces directional error by 64% and magnitude error by 68%, underscoring the importance of spatial acoustic sensing in resolving slip direction ambiguity. We further evaluate A-SLIP in closed-loop reactive control, and find that it enables reliable and low-cost real-time estimation of in-hand slip. Project videos and additional details are available at a-slip.github.io.**

I. Introduction

Reliable in-hand manipulation requires a robot to maintain stable and controlled contact with grasped objects throughout a task. A central challenge is real-time detection and estimation of slip, defined as relative motion between the gripper fingers and the surface of the object. When a slip goes unobserved or uncorrected, it can lead to dropping objects, task failures, or unintended disturbances to the environment. Detecting slip is particularly challenging because it is transient, directionally varying, and often occluded by the gripper. Moreover, slip events frequently occur on timescales faster than a purely reactive control loop can compensate without advanced sensing. As a result, reliable slip estimation is a critical capability for robust in-hand manipulation and the deployment of robot manipulators in unstructured real-world environments.

Researchers have approached in-hand slip sensing through various modalities. Wrist-mounted force-torque sensors can detect the onset of slip through observing changes in the measured wrench, but they provide ambiguous signals for slip direction and are sensitive to external contact disturbances unrelated to the slip. Capacitive and resistive tactile sensor arrays offer spatially resolved pressure measurements, and prior work has demonstrated their ability to infer slip from shear and pressure redistribution patterns [1]. However, these sensors are sensitive to wear, require complex fabrication, and degrade over repeated use due to difficult-to-model soft sensor phenomena such as creep and hysteresis. Vision-based tactile sensors such as GelSight [2] and DIGIT [3] embed cameras beneath a deformable gel surface and can capture rich contact geometry with high spatial resolution. However, vision-based tactile sensors commonly suffer from bulky form factors, low data-acquisition rates, limited scalability and sensor coverage, and low durability under repeated contact due to a thin spectral coating, restricting their practical deployment.

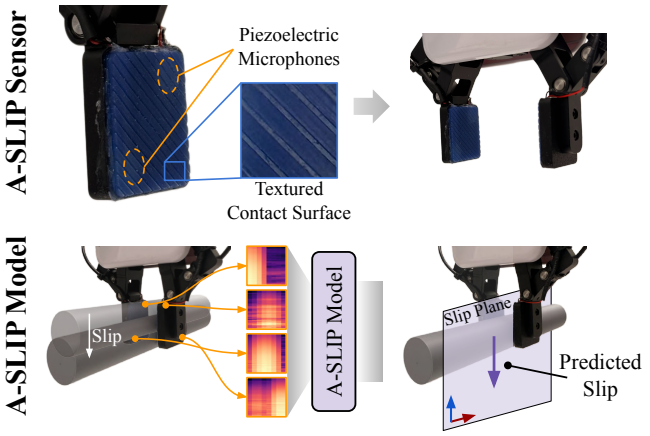


Fig. 1: **Overview of A-SLIP:** Piezoelectric microphones embedded behind textured silicone contact pads capture structure-borne vibrations during slip. Multi-channel log-mel spectrograms are processed by a convolutional network with channel and temporal attention to jointly estimate slip presence, magnitude, and direction as $\mathbf{v}_t \in \mathbb{R}^2$.

A promising alternative is acoustic sensing. When a slip occurs at the gripper-object interface, friction and surface asperities generate structure-borne vibrations that propagate through the sensor body and can be captured by microphones embedded in the fingers. Prior work has shown that acoustic signals carry information about contact events and surface properties during manipulation [4]–[6], and that piezoelectric microphones can detect the onset of slip [7] in the presence of robot operating noise. However, existing acoustic approaches have largely been limited to binary slip detection and do not address the estimation of slip direction or magnitude, which are necessary for closed-loop grasp correction. Moreover, prior work has generally not explored multi-channel acoustic fusion or learning-based approaches for continuous slip vector estimation.

In this work, we present Acoustic Sensing for Learning In-hand slip Parameters (A-SLIP), an embedded acoustic

[1]Robotics Institute, Carnegie Mellon University, USA; [*]Equal contribution.

sensing system for slip direction and magnitude estimation. We design a low-profile gripper sensor consisting of a textured silicone contact surface with embedded piezoelectric microphones. The textured silicone promotes structured vibrations at the contact interface during slip, while the piezoelectric microphones provide broadband sensitivity to structure-borne sound with minimal footprint. Compared to vision-based tactile sensors, the A-SLIP design requires no optics, illumination, or cameras, resulting in a sensor that is more compact, durable, and low cost. Building on this hardware, A-SLIP learns to map synchronized multi-microphone spectrograms to a continuous planar slip vector that jointly encodes slip event, direction, and magnitude within a unified prediction framework.

The contributions of this work are as follows:

- A-SLIP sensor, a low-profile and low-cost acoustic gripper sensor system based on a textured silicone contact surface with embedded piezoelectric microphones that is durable and suitable for real-world deployment across parallel-jaw gripper platforms.
- A-SLIP model, a slip prediction network that jointly estimates slip presence, magnitude, and direction through a unified multi-objective formulation with a two-stage pre-training and finetuning strategy.
- Experiments with A-SLIP on real-time estimation of in-hand slip and reactive closed-loop control.

II. Related Work

Prior work explored three relevant directions: tactile sensing hardware, slip estimation methods, and acoustic sensing as an emerging modality for contact-rich robotics.

A. Tactile Sensing

Touch is a fundamental modality underlying human dexterity, and tactile sensing has been extensively studied to endow robots with similar capabilities. Force-torque sensors characterize touch by measuring contact forces and moments directly [8]–[11]; vision-based tactile sensors infer touch from high-resolution surface deformation [3], [12]–[14]; capacitive and magnetic-based sensors localize contact by measuring changes in electric or magnetic fields induced by deformation or proximity [15]–[18]. Multimodal tactile sensors combine sensing modalities to extract more comprehensive information from physical interactions [19], [20]. Prior works have demonstrated that tactile sensing improves manipulation performance in various domains, such as object geometry recovery [21], object material property estimation [22], and dexterous in-hand manipulation [23].

However, existing sensors often remain difficult to deploy in general-purpose manipulation due to form factor, calibration complexity, and cost. In contrast, we propose acoustic sensing, which characterizes contact through structure-borne mechanical vibrations generated by physical interaction. This tactile sensing modality is compact, low-cost, and capable of delivering high-frequency, low-latency signals suitable for real-time control.

B. Slip Estimation

Slip estimation has been approached through sensing modalities with distinct trade-offs in signal richness, latency, and deployability. Wrist-mounted force-torque sensors can detect slip onset via abrupt changes in the measured wrench [24], [25], but provide limited directional information and often conflate slip-induced loads with external contact disturbances [1]. Tactile sensor arrays enable spatially resolved estimation by tracking shear and pressure redistribution across the contact patch [26], while vision-based tactile sensors such as GelSight [2] and DIGIT [3] further extend this capability to high-resolution contact geometry reconstruction. Learning-based methods, including convolutional and recurrent architectures for slip classification from tactile sequences [27] and self-supervised approaches that reduce labeling requirements [28], have improved prediction from raw tactile streams. However, these approaches often rely on vision-based tactile sensors with thin compliant gel surfaces and optical coatings that are susceptible to wear, hysteresis, and performance degradation under repeated shear [3]. As a result, many datasets emphasize contact events or controlled interactions rather than sustained slip.

Acoustic sensing offers a compelling alternative for slip estimation. Piezoelectric microphones are rigid, wear-resistant, and can fit into smaller form factors, while slip-induced vibrations propagate through the gripper on timescales faster than vision-based feedback can resolve [29]. Prior acoustic approaches have largely focused on binary slip detection [30] or contact event recognition [31] and do not estimate slip direction or magnitude. A-SLIP addresses these limitations through a multi-channel acoustic sensing pipeline and a learning-based architecture for continuous planar slip vector estimation, enabling actionable feedback for closed-loop manipulation.

C. Acoustic Sensing for Manipulation

Acoustic sensing in robotics can be categorized along two orthogonal dimensions. The first is the *propagation medium*: airborne acoustics captures sounds transmitted through air, typically corresponding to human-audible interaction cues such as pouring [32], whereas structure-borne acoustics captures mechanical vibrations transmitted through rigid bodies and often encodes contact phenomena that are imperceptible without instrumentation. The second is the sensing mode: *passive sensing* listens to naturally occurring signals during interaction [5], [33], while *active sensing* emits a probing signal and analyzes the response [34], [35]. Prior work has leveraged acoustic sensing in manipulation both to learn task-relevant representations, such as material properties or contact states, for downstream performance [31], and as an auxiliary modality within end-to-end learning pipelines [36].

Slip events generate brief, contact-localized, high-frequency structure-borne vibrations that encode both the presence and direction of relative motion at the contact interface. A-SLIP leverages passive structure-borne acoustic sensing with piezoelectric microphones embedded in the gripper and learns directional asymmetries in slip-induced

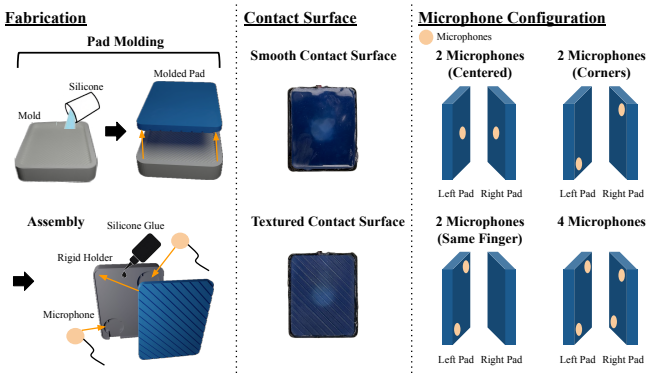


Fig. 2: **Design of the A-SLIP Sensor.** (*Left*) Fabrication pipeline: silicone is cast in a 3D-printed mold to form a compliant contact pad, which is bonded to a rigid holder with embedded piezoelectric microphones. (*Center*) Contact surface variants: smooth and textured silicone pads; the textured surface introduces controlled asperities that produce more directionally informative vibrations during slip. (*Right*) Evaluated microphone placements: three two-microphone layouts (centered, corners, and same-finger) and a four-microphone layout with two microphones per finger to increase contact-region coverage.

vibrations to estimate a continuous planar slip vector encoding both slip direction and magnitude.

III. Problem Formulation

The problem is to infer in-hand planar object slip presence, direction, and magnitude from acoustic observations during grasped manipulation. Let $\mathbf{X}_t^n = \{\mathbf{x}_t^1, \mathbf{x}_t^2, ..., \mathbf{x}_t^n\}$ denote synchronized audio measurements of n microphones embedded in the gripper at time t, where $\mathbf{x}_t^i \in \mathbb{R}^{T \times F}$ represents a short-time spectral representation over a temporal window of duration T. These measurements provide indirect observations of vibrations at the gripper-object interface.

We define the slip state at time t as a planar slip vector

$$\mathbf{v}_t = (v_x, v_z) \in \mathbb{R}^2,$$

where the direction of $\mathbf{v}_t$ encodes the instantaneous direction of slip in the gripper plane and its magnitude $\|\mathbf{v}_t\|$ corresponds to the intensity of slip. The no-slip condition is captured naturally by $\mathbf{v}_t = \mathbf{0}$.

The goal is to learn a function

$$f_\theta : \mathbf{X}_t^n \mapsto \mathbf{v}_t,$$

parameterized by θ. As $\mathbf{X}_t^n$ depends on the sensor design and microphone placement, we also seek a sensor configuration to minimize slip-estimation error.

IV. Methods

A-SLIP achieves state-of-the-art acoustic in-hand slip estimation through a unique recipe of hardware design, network architecture, and dataset curation.

A. Hardware Design

We propose a low-profile acoustic gripper sensor design composed of two primary components: a molded silicone contact pad and a rigid finger-mounted holder with embedded piezoelectric microphones (Fig. 2).

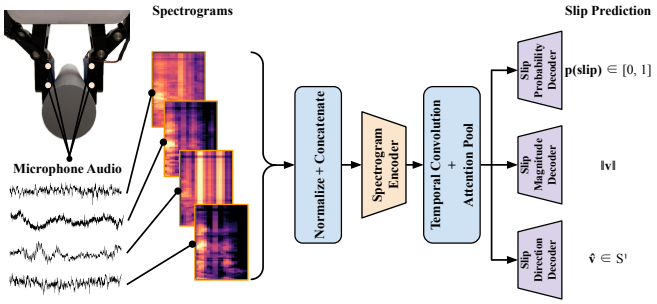


Fig. 3: **A-SLIP Model Architecture.** Log-mel spectrograms from synchronized microphones are normalized and fused via a learned channel attention module. The fused representation passes through 2D convolutional layers preserving temporal resolution, then 1D temporal convolutions modeling slip dynamics. A temporal attention pooling module aggregates features into a latent vector passed to three heads: a *slip classification head* predicting $p(\text{slip}) \in [0, 1]$, a *magnitude regression head* predicting $\|\mathbf{v}\|$, and a *direction head* predicting $\hat{\mathbf{v}} \in \mathcal{S}^1$.

To promote conformal contact with grasped objects while remaining sufficiently stiff to efficiently transmit structure-borne vibrations to the embedded microphone, we fabricate the contact pad with a two-part platinum-cure liquid silicone rubber (Shore 30A). We mix the silicone at a 1:1 volume ratio and cast it with a custom 3D-printed mold. After curing, we demold the pad and bond it to the rigid holder with silicone adhesive to ensure strong mechanical and acoustic coupling between the contact surface and sensor body.

Inspired by prior work on vibration-inducing surface textures for other tactile modalities [37], we design two mold variants to produce either a contact face with imprinted regular textures or a smooth contact surface (Fig. 2, center). In experiments the textured pad shows a 62.90 % reduction in directional MAE compared to the smooth pad, suggesting that the textured surface introduces controlled asperities that modulate contact-induced vibrations and produce richer, more directionally informative acoustic signatures.

Each rigid finger-mounted holder houses one or more microphones positioned flush with the holder's top surface. After bonding the silicone pad with the holder, the microphones sit directly beneath the pad, maximizing sensitivity to contact-induced structure-borne vibrations. In experiments (Sec. V-A), we evaluate four microphone configurations (Fig. 2, right) corresponding to different holder designs.

We mount the assembled sensor on both fingers of a parallel-jaw gripper, replacing the default contact surfaces. The rigid holders match the gripper finger mounting geometry, allowing drop-in installation without structural modification. Microphone signals route through thin-gauge wires along the finger body to an external data acquisition mixer that synchronizes multi-channel audio captured at a fixed sampling rate. The resulting sensor adds minimal bulk to the gripper profile and preserves workspace clearance.

B. Slip Prediction Model

Slip events manifest as brief broadband friction-induced vibrations at the contact interface. To capture both the

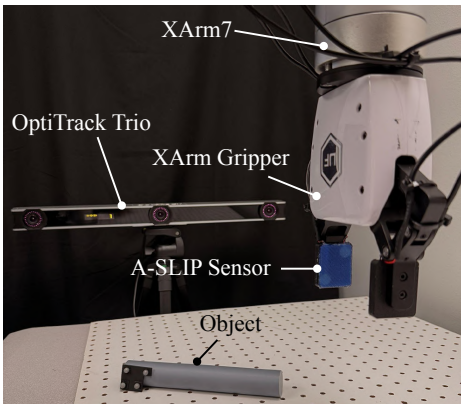


Fig. 4: **A-SLIP System.** We mount A-SLIP sensors on an XArm gripper attached to an XArm7 robot. To obtain ground-truth in-hand slip, we use an OptiTrack Trio to track poses of the left finger and the object, each with reflective markers attached.

frequency content and temporal evolution of these vibrations, we represent each microphone signal as a log-mel spectrogram computed over 200 ms windows. Each input sample is a tensor $\mathbf{X} \in \mathbb{R}^{n \times M \times T}$, where the n channels correspond to the n microphones, M is the number of mel-frequency bins, and T is the number of time frames. Spectrograms are normalized using dataset-level mean and variance statistics to reduce sensitivity to gain and contact variability.

A-SLIP's slip prediction network is a convolutional architecture designed to preserve fine-grained temporal cues critical for slip direction estimation. A learnable channel attention mechanism first fuses the multi-microphone streams: frequency-averaged spectrograms are passed through a lightweight temporal convolutional gating network that predicts per-channel weights, allowing the model to emphasize microphone with stronger slip-related cues. The fused representation is then processed by a stack of 2D convolutional layers interleaved with batch normalization, ReLU activations, dropout, and frequency-only max pooling to preserve temporal resolution. Subsequent 1D temporal convolution layers capture short-term dynamics associated with slip onset and direction changes. Finally, a learned temporal attention pooling mechanism aggregates features into a fixed-length latent vector, which is passed to three prediction heads: a slip classification head outputting $p(\text{slip})$, a magnitude regression head, and a direction head predicting a unit-normalized 2D vector as shown in Fig. 3.

We train the network with a multi-objective loss that jointly supervises slip detection, magnitude estimation, and direction estimation. Let $\mathbf{v}^* = (v_x^*, v_z^*) \in \mathbb{R}^2$ denote the ground-truth slip vector defined in the slip plane of the parallel jaw gripper and $\hat{\mathbf{v}} \in \mathbb{R}^2$ the predicted vector. We supervise slip presence with a binary cross-entropy loss

$$\mathcal{L}_{\text{slip}} = \text{BCE}\left(p(\text{slip}),\, \mathbf{1}\left[\|\mathbf{v}^*\| > \epsilon\right]\right),$$

where ϵ defines the slip magnitude threshold to label a sample as in slip.

We supervise slip magnitude with a Huber loss applied only on frames labeled as slip,

$$\mathcal{L}_{\text{mag}} = \text{Huber}\left(\|\hat{\mathbf{v}}\|,\, \|\mathbf{v}^*\|\right).$$

We supervise slip direction with a cosine similarity loss,

$$\mathcal{L}_{\text{dir}} = 1 - \hat{\mathbf{d}}^\top \mathbf{d}^*,$$

where $\hat{\mathbf{d}} = \hat{\mathbf{v}}/\|\hat{\mathbf{v}}\|$ and $\mathbf{d}^* = \mathbf{v}^*/\|\mathbf{v}^*\|$ are the predicted and ground-truth unit direction vectors, respectively. To mitigate vanishing gradients when the predicted direction opposes the ground truth, we additionally apply an auxiliary loss on the unnormalized direction logits $\hat{\mathbf{v}}$. We further include a temporal smoothness regularizer that penalizes large angular deviations between consecutive predicted directions,

$$\mathcal{L}_{\text{smooth}} = 1 - \hat{\mathbf{d}}_t^\top \mathbf{d}_{t-1}^*,$$

and define the final training objective as a weighted sum of all components, $\mathcal{L} = \lambda_{\text{slip}}\mathcal{L}_{\text{slip}} + \lambda_{\text{mag}}\mathcal{L}_{\text{mag}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}$. In our experiments, we set $\lambda_{\text{dir}} = 2.0$ to prioritize direction estimation, $\lambda_{\text{slip}} = 1.0$, $\lambda_{\text{mag}} = 0.5$, and $\lambda_{\text{smooth}} = 0.1$, treating the smoothness term as a light regularizer.

C. Data Collection and Model Training

Learning slip direction and magnitude from audio can require both large amounts of labeled data and accurate ground-truth supervision. Since collecting large-scale datasets with precise slip labels during real manipulation needs specialized external tracking systems, we adopt a two-stage data collection and training strategy that separates representation learning from task-specific adaptation.

We mount the parallel-jaw gripper equipped with the A-SLIP sensor on a robot arm and use a motion capture system calibrated to the gripper's grasp plane to precisely track the in-plane motion of both the gripper and the object, allowing slip to be inferred from their relative poses (Fig. 4). First, we collect an audio dataset of *robot-induced slip* by executing randomized robot motions that sweep the gripper across a stationary calibrated 3D printed probe (Fig. 5, "Robot-Induced Slip"). For this dataset, we compute slip direction and magnitude labels directly from the recorded robot state. Second, we collect a small dataset of *externally-induced slip* by manually perturbing five objects (four from the YCB dataset [38]) grasped by a stationary gripper (Fig. 5, "Externally-Induced Slip"). For each object, we record 30-second trials while using a motion capture system to automatically obtain slip direction and magnitude labels. We perform 30 trials with the robot on and stationary, 10 with the robot executing random motions, and 20 with the robot off.

We use the *robot-induced slip* dataset to pretrain the model to learn a general acoustic representation of slip. We then finetune the pretrained model on the *externally-induced slip* dataset to adapt to the target sensing scenario, where slip arises from external disturbances. During finetuning, we freeze the audio encoder and optimize only the task-specific

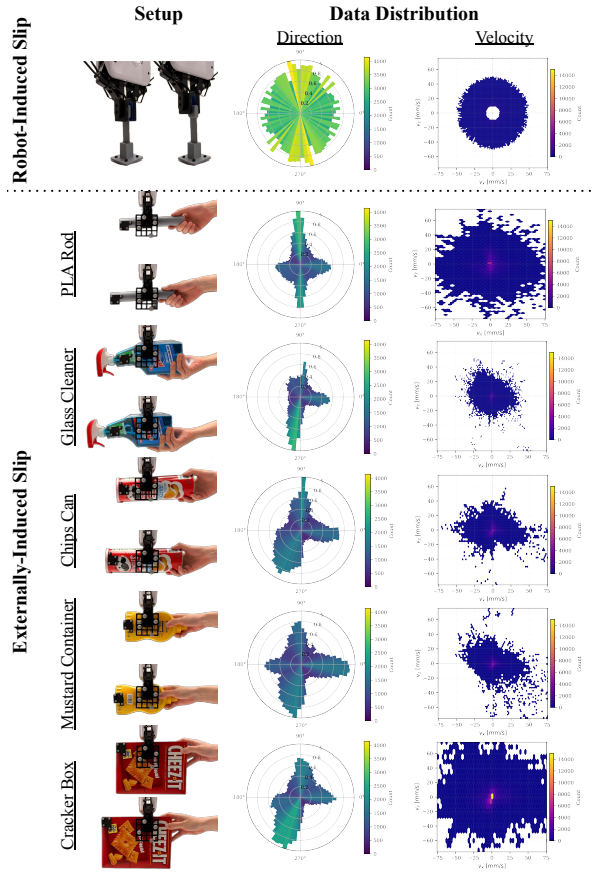


Fig. 5: **A-SLIP Dataset Distribution.** (*Top*) Robot-induced slip data collected automatically via randomized robot motions sweeping the gripper across a stationary probe, with labels derived from robot states. (*Bottom*) Externally-induced slip data collected by manually perturbing grasped objects, with labels from OptiTrack-tracked rigid-body motion. Counts indicate the number of audio slices in the dataset.

prediction heads. This preserves the learned acoustic representation while enabling specialization for slip detection, magnitude estimation, and direction prediction.

We train the model using the Adam optimizer with a learning rate of 10^{-3} and weight decay of 10^{-4}. To improve robustness across surface textures and contact conditions, we apply SpecAugment-style time and frequency masking [39] as well as random gain augmentation. To mitigate class imbalance, we subsample and reweight frames without slip. We train the models for up to 1000 epochs, stopping early based on validation loss. Inference treats predictions with $p(\text{slip}) < 0.5$ as a zero slip vector.

V. Evaluation

We evaluate A-SLIP against baselines and system ablations to isolate contributions of the training regime, microphone number and placement, and spectrogram temporal window size. Additionally, we evaluate robustness to robot operating noise, and system integration into closed-loop reactive control tasks. All experiments use a parallel-jaw gripper with A-SLIP sensors and objects tracked by an external motion capture system to provide the ground truth.

TABLE I: Slip prediction model comparison. Binary slip accuracy (Det.). Dir. MAE: Slip direction MAE (Dir. MAE). Slip magnitude RMSE (Mag. RMSE).

Method	**Det.** (%)	**Dir. MAE** (deg)	**Mag. RMSE** (mm)
SVM	73.0	19.2 ± 28.6	1.9 ± 2.4
Single mic (no pretrain)	43.3	28.5 ± 26.5	2.0 ± 1.0
Single mic (pretrain + finetune)	70.2	20.7 ± 22.4	2.7 ± 0.9
Pretrain only (2-mic, centered)	60.4	29.3 ± 21.5	1.5 ± 1.2
Pretrain only (2-mic, corners)	63.1	26.8 ± 21.1	1.5 ± 1.2
Pretrain only (2-mic, same finger)	61.2	26.7 ± 21.2	1.6 ± 1.2
Pretrain only (4-mic)	72.4	21.6 ± 8.0	3.2 ± 1.4
A-SLIP (2-mic, centered, finetuned)	63.6	20.4 ± 16.0	1.0 ± 1.3
A-SLIP (2-mic, corners, finetuned)	71.9	18.1 ± 9.3	0.6 ± 0.6
A-SLIP (2-mic, same finger, finetuned)	72.1	19.0 ± 10.1	0.7 ± 0.5
A-SLIP (4-mic, finetuned)	**81.8**	$\mathbf{14.1 \pm 6.9}$	1.0 ± 0.9
A-SLIP (100 ms window)	72.5	$\mathbf{8.2 \pm 9.6}$	$\mathbf{0.6 \pm 0.7}$
A-SLIP (200 ms window)	81.8	14.1 ± 6.9	1.0 ± 0.9
A-SLIP (300 ms window)	**88.5**	20.9 ± 6.5	1.4 ± 1.1

A. Baselines

We compare against baselines from prior work and against several system ablations. Prior studies on embedding low-profile microphones into end-effectors commonly use SVM regressors [6] or single-microphone sensing for event detection [40]. In Table I, we compare the A-SLIP model against an SVM baseline and against single-microphone variants trained with and without pretraining. A-SLIP model achieves the best overall performance, improving detection accuracy by up to 12% and reducing directional MAE by up to 32% relative to these baselines. Compared specifically to the single-microphone variants, the 4-microphone finetuned model reduces directional error by 64% and magnitude RMSE by 68%, showing that the gains come from both combining A-SLIP's training procedures with spatially distributed multi-channel sensing.

Additionally, we compare four microphone configurations in Fig. 2. Among the 2-microphone variants, microphone placement influences performance by affecting how well the sensor captures vibration asymmetries across the grasp. A-SLIP (2-mic, corners, finetuned) achieves the best performance among the 2-microphone configurations, reducing directional MAE by approximately 5% relative to A-SLIP (2-mic, same finger, finetuned) and by about 13% relative to A-SLIP (2-mic, centered, finetuned). This suggests that distributing microphones across opposite fingers helps preserve bilateral differences in vibration propagation that encode slip direction. In contrast, when both microphones are placed on a single finger, the model cannot directly observe cross-finger vibration differences, which weakens the directional signal available for inference. Even with bilateral sensing, the performance gap between the corners and centered placements indicates that increasing the spatial baseline between microphones further improves sensitivity to directional vibration patterns. Expanding to A-SLIP (4-mic, finetuned) provides an additional 22% reduction in directional MAE relative to the best 2-microphone configuration and improves slip detection accuracy by 14%. These gains suggest that denser

TABLE II: Per-object slip prediction results. Binary slip accuracy (Det.). Slip direction MAE (Dir. MAE). Slip magnitude RMSE (Mag. RMSE).

Object	Train	Det. (%)	Dir. MAE (deg)	Mag. RMSE (mm)
PLA Rod	Per-obj.	94.5	26.07 ± 9.8	$\mathbf{2.44 \pm 1.20}$
	All-obj.	**94.7**	$\mathbf{25.58 \pm 9.1}$	2.72 ± 1.30
Glass Cleaner	Per-obj.	**81.1**	15.17 ± 7.4	0.60 ± 0.72
	All-obj.	78.7	$\mathbf{14.38 \pm 6.8}$	$\mathbf{0.58 \pm 0.68}$
Chips Can	Per-obj.	**71.8**	9.01 ± 6.1	$\mathbf{0.40 \pm 0.63}$
	All-obj.	71.0	$\mathbf{8.58 \pm 6.5}$	0.41 ± 0.66
Mustard Container	Per-obj.	**84.0**	24.14 ± 9.6	0.62 ± 0.82
	All-obj.	83.7	$\mathbf{22.20 \pm 8.9}$	$\mathbf{0.60 \pm 0.80}$
Cracker Box	Per-obj.	79.4	7.61 ± 5.9	0.62 ± 0.74
	All-obj.	**80.6**	$\mathbf{6.90 \pm 5.5}$	$\mathbf{0.55 \pm 0.70}$

spatial sampling of the vibration field allows the model to learn more stable cross-channel relationships associated with slip direction, while magnitude estimation appears largely governed by overall vibration energy and therefore benefits less from additional channels.

The bottom rows of Table I evaluate the effect of temporal window size by comparing spectrogram windows of 100 ms, 200 ms, and 300 ms. Increasing the window size consistently improves slip detection accuracy, indicating that longer temporal context makes it easier for the model to distinguish sustained slip events from transient contact noise. At the same time, both directional and magnitude estimation degrade as the window grows longer. A likely explanation is that slip direction and intensity often vary within a single window, particularly during externally induced disturbances, and aggregating over longer time intervals blurs these instantaneous dynamics. Shorter windows better preserve the local structure of the slip signal but provide less evidence for reliably detecting whether slip is occurring. The 200 ms window represents a balance between these effects.

B. Cross-Object Generalization

Table II reports per-object results using the 4-microphone configuration to assess sensitivity to object geometry and surface material. We compare models trained on each object individually (per-obj.) against a single model trained on all objects jointly (all-obj.). The joint model achieves comparable or improved directional accuracy across all objects, reducing directional MAE by approximately 2%–9% relative to the per-object specialist models for Glass Cleaner, Chips Can, Mustard Container, and Cracker Box, while maintaining nearly identical detection accuracy. These results suggest that training across diverse contact surfaces improves the robustness of the learned acoustic representation without sacrificing object-specific performance. Magnitude RMSE remains largely unchanged across objects and training regimes, indicating that slip magnitude estimation is relatively invariant compared to directional estimation.

C. Slip with Robot Noise

A concern for acoustic sensing is whether vibrations generated by the robot itself interfere with the slip signal.

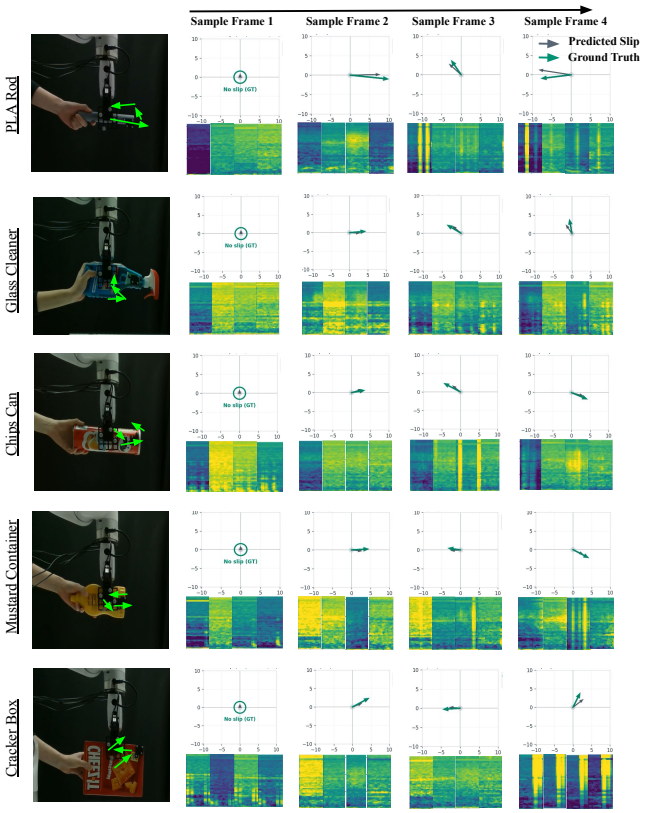


Fig. 6: **Qualitative Evaluation of A-SLIP.** Each row shows a different object; each column shows a sample evaluation frame with predicted (gray) and ground-truth (green) slip vectors overlaid on the contact image alongside per-channel log-mel spectrograms. A-SLIP accurately estimates slip direction and magnitude across objects with varying geometry and surface material, even under impulsive externally induced slip.

To isolate this effect, we evaluate each *finetuned* model on the robot-induced pretraining validation set, where the robot actively executes the slip motion and the audio contains both slip-induced vibrations and robot operating noise. Robot noise does not substantially degrade performance for the 4-microphone model. It achieves a directional MAE of 15.9 ± 16.6 degrees and a magnitude RMSE of 0.5 ± 0.2 mm. Compared with its externally induced performance, the directional error increases by only 12.8%, suggesting that robot operating sound is not the dominant source of error for the best-performing model. The relative ordering across microphone layouts is also consistent. Among the 2-microphone variants, the centered layout is most sensitive to robot noise, reaching 39.2 ± 26.0 degrees directional MAE and 0.7 ± 0.4 mm magnitude RMSE, corresponding to a 92.2% increase in directional error relative to the externally induced slip setting. In contrast, the corners layout reaches 21.3 ± 20.9 degrees and 0.5 ± 0.4 mm, only a 17.7% increase in directional error, while the same-finger layout achieves 17.4 ± 18.7 degrees and 0.6 ± 0.3 mm, an 8.4% reduction in directional MAE. These results suggest that configurations with spatial coverage remain accurate under active robot motion, whereas the centered 2-microphone layout degrades.

TABLE III: Task performances. Success rate (Succ.), stopping error (Δx), and RMSE of gripper-object pose along trajectory (Pose RMSE).

	Task 1: Slip-Stop		**Task 2: Slip-Track**	
	Succ., Δx (mm)		**Pose RMSE** (mm)	
Object	SVM	A-SLIP	SVM	A-SLIP
PLA Rod	7/10, 39.7 ± 38.7	**10/10**, $\mathbf{16.7 \pm 8.9}$	32.4 ± 20.2	$\mathbf{16.9 \pm 12.4}$
Glass Cleaner	6/10, 13.4 ± 18.7	**10/10**, $\mathbf{10.5 \pm 5.7}$	32.4 ± 18.1	$\mathbf{15.0 \pm 9.5}$
Chips Can	8/10, $\mathbf{16.7 \pm 16.4}$	**10/10**, 19.8 ± 6.6	33.8 ± 21.7	$\mathbf{12.9 \pm 9.4}$
Mustard Container	5/10, 17.4 ± 31.3	**10/10**, $\mathbf{9.7 \pm 6.7}$	25.0 ± 14.4	$\mathbf{20.5 \pm 14.4}$
Cracker Box	5/10, 44.3 ± 27.6	**10/10**, $\mathbf{15.8 \pm 7.1}$	35.5 ± 20.4	$\mathbf{13.0 \pm 9.3}$
Overall	31/50, 26.3 ± 29.6	**50/50**, 14.5 ± 7.8	32.0 ± 19.3	$\mathbf{15.8 \pm 11.3}$

D. Reactive Control

We evaluate A-SLIP in two closed-loop control tasks, where real-time audio is streamed to the model and predictions directly drive robot motion.

In the first task (Fig. 7, left), the robot pushes a grasped object against a wall and stops upon detecting in-hand slip. We compare the A-SLIP model with the SVM baseline over 10 trials per object. In each trial, the robot begins with the object in contact with the wall, retracts 10 cm, and re-approaches. We measure the stopping error along the motion direction between the robot's initial wall-contact pose and final stopping pose, denoted as Δx, and we consider a trial successful only if the object makes contact with the wall and the robot stops before pushing it past its full length (Table III, left). Overall, A-SLIP achieves a 100 % success rate, while the SVM baseline succeeds in 62 % of the trials and yields a mean stopping error 81.4 % larger than A-SLIP. The SVM baseline also shows higher variance, largely driven by failure cases, and varies significantly across objects, indicating limited robustness to different contact conditions. These results suggest that A-SLIP provides more reliable slip detection in closed-loop control, reducing both missed slip events and unstable behavior caused by incorrect predictions.

In the second task (Fig. 7, right), as an experimenter induces slip, the robot continuously tracks the predicted slip vector to maintain a stable object-gripper relative pose. We perform 10 trials for each object. Qualitatively, A-SLIP enables the robot to reliably follow the object as it moves, whereas the SVM baseline often fails to produce meaningful motion due to inaccurate slip predictions. Quantitatively, we evaluate how well the gripper maintains a constant relative pose with the object in the grasp plane by computing the RMSE of gripper-object poses along the trajectory (Table III, right). Overall, A-SLIP achieves an RMSE that is 50.5 % lower than the SVM baseline. These results indicate that A-SLIP predicts slip direction and magnitude with sufficient accuracy to enable fast and reliable robot reactions to in-hand slip, supporting real-time feedback-based tracking.

VI. Limitations and Future Work

A-SLIP has several limitations that suggest directions for future work. The system estimates planar slip only and does not model rotational slip about the grasp axis; extending the slip representation to include rotational components would provide more complete coverage of in-hand motion.

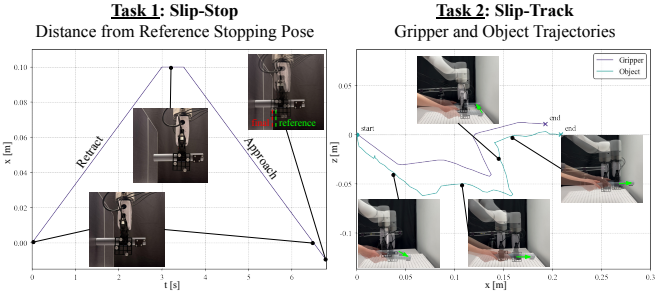


Fig. 7: **Reactive Control.** A-SLIP predicts slip direction and magnitude in real time, enabling rapid robot responses to in-hand slip. (*Left*) Task 1: the robot pushes an object against a wall and stops automatically upon in-hand slip detection. (*Right*) Task 2: as an experimenter induces slip, the robot follows the model-predicted slip vector to maintain a stable grasp.

Although the textured silicone pad promotes structured vibrations, acoustic signatures vary with object surface material and fully smooth or compliant objects may produce weaker signals and degrade accuracy, motivating domain adaptation or online recalibration strategies. The finetuning stage relies on motion capture for ground-truth labels, which may be unavailable in many settings; self-supervised or weakly supervised labeling would reduce this dependency. Evaluation is limited to a single parallel-jaw gripper, and differences in finger geometry or material may require adaptation. The 200 ms inference window introduces latency that could limit performance in high-speed tasks, suggesting further exploration of shorter windows with history of observations or causal streaming architectures.

VII. Conclusion

We present A-SLIP, an acoustic sensing system for continuous planar slip vector estimation in robotic in-hand manipulation. By embedding low-cost piezoelectric microphones behind textured silicone contact pads on a parallel-jaw gripper, A-SLIP captures structure-borne vibrations induced by gripper-object slip without requiring cameras, optics, or complex fabrication. Our slip prediction network processes synchronized multi-channel log-mel spectrograms using a convolutional architecture with learned channel attention and temporal attention pooling, jointly estimating slip presence, magnitude, and direction within a unified multi-objective framework. A two-stage training strategy that combines pretraining on robot-induced slip data with finetuning enables the model to learn transferable acoustic slip representations and adapt them to task-specific conditions.

Experimental results show that A-SLIP achieves strong performance in slip detection, direction estimation, and magnitude regression. In particular, the finetuned 4-microphone configuration outperforms all baselines, including the SVM baseline and pretraining-only variants, across all evaluation metrics. These results demonstrate that multi-channel acoustic sensing, when combined with learning-based fusion, provides a practical and effective solution for continuous slip vector estimation required for closed-loop grasp correction. Overall, A-SLIP strengthens acoustic sensing as a compelling

modality for slip estimation, offering advantages in form factor, durability, and deployment cost.

ACKNOWLEDGMENTS

This work was supported by Samsung Research America and NSF Graduate Research Fellowship under Grant No. DGE2140739.

REFERENCES

[1] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson, "Measurement of shear and slip with a gelsight tactile sensor," in *2015 IEEE international conference on robotics and automation (ICRA)*, pp. 304–311, IEEE, 2015.

[2] 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.

[3] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[4] Y. Mao, U. Yoo, Y. Yao, S. N. Syed, L. Bondi, J. Francis, J. Oh, and J. Ichnowski, "Visuo-acoustic hand pose and contact estimation," *arXiv preprint arXiv:2508.00852*, 2025.

[5] M. Lee, U. Yoo, J. Oh, J. Ichnowski, G. Kantor, and O. Kroemer, "Sonicboom: Contact localization using array of microphones," *IEEE Robotics and Automation Letters*, 2025.

[6] G. Zöller, V. Wall, and O. Brock, "Active acoustic contact sensing for soft pneumatic actuators," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7966–7972, IEEE, 2020.

[7] Y. Mao, B. P. Duisterhof, M. Lee, and J. Ichnowski, "Hearing the slide: Acoustic-guided constraint learning for fast non-prehensile transport," 2025.

[8] M. Y. Cao, S. Laws, and F. R. y Baena, "Six-axis force/torque sensors for robotics applications: A review," *IEEE Sensors Journal*, vol. 21, no. 24, pp. 27238–27251, 2021.

[9] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim, "Supervised autonomous robotic soft tissue surgery," *Science translational medicine*, vol. 8, no. 337, 2016.

[10] P. Nadeau, M. Giamou, and J. Kelly, "Fast object inertial parameter identification for collaborative robots," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 3560–3566, IEEE, 2022.

[11] S. Suresh, M. Bauza, K.-T. Yu, J. G. Mangelson, A. Rodriguez, and M. Kaess, "Tactile slam: Real-time inference of shape and pose from planar pushing," in *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 11322–11328, IEEE, 2021.

[12] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021.

[13] A. Alspach, K. Hashimoto, N. Kuppuswamy, and R. Tedrake, "Soft-bubble: A highly compliant dense geometry tactile sensor for robot manipulation," in *2019 2nd IEEE International Conference on Soft Robotics (RoboSoft)*, pp. 597–604, IEEE, 2019.

[14] M. Oller, M. P. i Lisbona, D. Berenson, and N. Fazeli, "Manipulation via membranes: High-resolution and highly deformable tactile sensing and control," in *Conference on Robot Learning*, pp. 1850–1859, PMLR, 2023.

[15] T. Hellebrekers, N. Chang, K. Chin, M. J. Ford, O. Kroemer, and C. Majidi, "Soft magnetic tactile skin for continuous force and location estimation using neural networks," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3892–3898, 2020.

[16] Y. Wi, J. Yin, E. Xiang, A. Sharma, J. Malik, M. Mukadam, N. Fazeli, and T. Hellebrekers, "Tactalign: Human-to-robot policy transfer via tactile alignment," *arXiv preprint arXiv:2602.13579*, 2026.

[17] X. Liu, W. Yang, F. Meng, and T. Sun, "Material recognition using robotic hand with capacitive tactile sensor array and machine learning," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–9, 2024.

[18] T. Yao, X. Guo, C. Li, H. Qi, H. Lin, L. Liu, Y. Dai, L. Qu, Z. Huang, P. Liu, *et al.*, "Highly sensitive capacitive flexible 3d-force tactile sensors for robotic grasping and manipulation," *Journal of Physics D: Applied Physics*, vol. 53, no. 44, p. 445109, 2020.

[19] Q. Mao, Z. Liao, J. Yuan, and R. Zhu, "Multimodal tactile sensing fused with vision for dexterous robotic housekeeping," *Nature Communications*, vol. 15, no. 1, p. 6871, 2024.

[20] C. Higuera, A. Sharma, T. Fan, C. K. Bodduluri, B. Boots, M. Kaess, M. Lambeta, T. Wu, Z. Liu, F. R. Hogan, *et al.*, "Tactile beyond pixels: Multisensory touch representations for robot manipulation," in *Conference on Robot Learning*, pp. 105–123, PMLR, 2025.

[21] H.-J. Huang, M. A. Mirzaee, M. Kaess, and W. Yuan, "Gelslam: A real-time, high-fidelity, and robust 3d tactile slam system," *arXiv preprint arXiv:2508.15990*, 2025.

[22] J. Han, S. Yao, and K. Hauser, "Estimating high-resolution neural stiffness fields using visuotactile sensors," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2255–2261, IEEE, 2025.

[23] J. Wang, Y. Yuan, H. Che, H. Qi, Y. Ma, J. Malik, and X. Wang, "Lessons from learning to spin "pens"," in *Conference on Robot Learning*, pp. 3124–3138, PMLR, 2025.

[24] R. A. Romeo and L. Zollo, "Methods and sensors for slip detection in robotics: A survey," *IEEE Access*, vol. 8, pp. 73027–73050, 2020.

[25] M. Stachowsky, T. Hummel, M. Moussa, and H. A. Abdullah, "A slip detection and correction strategy for precision robot grasping," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 5, pp. 2214–2226, 2016.

[26] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing," in *Conference on Robot Learning*, pp. 2557–2578, PMLR, 2025.

[27] W. Yuan, S. Wang, S. Dong, and E. Adelson, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5580–5588, 2017.

[28] C. Higuera, A. Sharma, C. K. Bodduluri, T. Fan, P. Lancaster, M. Kalakrishnan, M. Kaess, B. Boots, M. Lambeta, T. Wu, *et al.*, "Sparsh: Self-supervised touch representations for vision-based tactile sensing," in *Conference on Robot Learning*, pp. 885–915, PMLR, 2025.

[29] S. Rangwala, F. Forouhar, and D. Dornfeld, "Application of acoustic emission sensing to slip detection in robot grippers," *International Journal of Machine Tools and Manufacture*, vol. 28, no. 3, pp. 207–215, 1988.

[30] W. Chen, H. Khamis, I. Birznieks, N. F. Lepora, and S. J. Redmond, "Tactile sensors for friction estimation and incipient slip detection—toward dexterous robotic manipulation: A review," *IEEE Sensors Journal*, vol. 18, no. 22, pp. 9049–9064, 2018.

[31] S. Lu and H. Culbertson, "Active acoustic sensing for robot manipulation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3161–3168, IEEE, 2023.

[32] H. Liang, S. Li, X. Ma, N. Hendrich, T. Gerkmann, F. Sun, and J. Zhang, "Making sense of audio vibration for liquid height estimation in robotic pouring," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, p. 5333–5339, IEEE, Nov. 2019.

[33] I. Andrussow, J. Solano, B. A. Richardson, G. Martius, and K. J. Kuchenbecker, "Adding internal audio sensing to internal vision enables human-like in-hand fabric recognition with soft robotic fingertips," in *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, pp. 01–08, IEEE, 2025.

[34] U. Yoo, Z. Lopez, J. Ichnowski, and J. Oh, "Poe: Acoustic soft robotic proprioception for omnidirectional end-effectors," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14980–14987, IEEE, 2024.

[35] K. Zhang, D.-G. Kim, E. T. Chang, H.-H. Liang, Z. He, K. Lampo, P. Wu, I. Kymissis, and M. Ciocarlie, "Vibecheck: Using active acoustic tactile sensing for contact-rich manipulation," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12278–12285, IEEE, 2025.

[36] 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.

[37] K. Dai, X. Wang, A. M. Rojas, E. Harber, Y. Tian, N. Paiva, J. Gnehm, E. Schindewolf, H. Choset, V. A. Webster-Wood, *et al.*, "Design of a biomimetic tactile sensor for material classification," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 10774–10780, IEEE, 2022.

[38] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks

for manipulation research," in *2015 international conference on advanced robotics (ICAR)*, pp. 510–517, IEEE, 2015.

[39] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[40] K. Zhang, C. Chang, S. Aggarwal, M. Veloso, F. Temel, and O. Kroemer, "Vibrotactile sensing for detecting misalignments in precision manufacturing," pp. 10408–10415, 10 2025.