SlipSense: Multimodal Tactile Learning for Low-Latency and Generalized Slip Detection

Anonymous Author(s)
Affiliation
Address
email

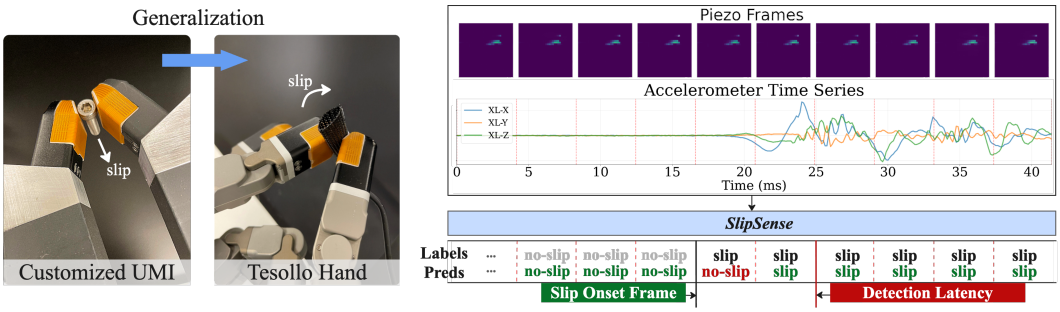


Figure 1: *Left:* TacV5 multimodal tactile sensor on UMI gripper and Tesollo hand, combining high-density pressure sensing (Piezo: 32×32, 240 Hz, 155 taxels/cm²) with high-bandwidth vibration sensing (XL: 3-axis MEMS accelerometer, 8 kHz). *Right:* We introduce *SlipSense*, a multimodal framework that fuses Piezo and XL signals for accurate, low-latency slip detection, achieving cross-object and cross-platform generalization.

Abstract: Slip detection is fundamental to dexterous manipulation, yet existing systems often lack precise detection latency characterization and cross-platform generalization. We present SlipSense, a multimodal tactile slip detection framework built on TacV5, a compact sensor integrating a 32×32 piezoresistive array (240 Hz) and a 3-axis MEMS accelerometer (8 kHz). The piezoresistive array captures spatial pressure distribution while the accelerometer captures friction-induced vibration, providing complementary slip cues. The framework performs modality-specific encoding, intra-sensor fusion, and cross-modal attention with causal temporal prediction at 240 Hz. Experiments on a 1.4M-frame dataset spanning 37 objects demonstrate that the two modalities are crucially complementary: it achieves 96.7% Macro F1 with false-positive rate below 1.6%, detecting 76% of slip events within 23.1 ms. Remarkably, trained purely on a UMI, SlipSense generalizes zero-shot to a Tesollo dexterous hand, transferring across unseen objects, distinct sensor units and new robotic platforms without any retraining.

Keywords: Slip Detection, Multimodal Tactile Sensing

1 Introduction

Slip detection is fundamental to dexterous manipulation, as slip is often one of the earliest observable indications of grasp instability. Early detection enables corrective action before object loss. In practice, slip rarely occurs abruptly; instead, it emerges progressively, spanning a continuum from localized partial motion (*incipient slip*) to full uncontrolled sliding (*gross slip*). Even subtle relative motion can destabilize a grasp and propagate into failure, making timely detection critical for robust manipulation. The entire slip process is therefore relevant to grasp stability, with particular emphasis on slip onset, referred to as the first occurrence of relative motion at the contact interface.

Submitted to the 10th Conference on Robot Learning (CoRL 2026). Do not distribute.

Slip generates diverse physical signatures at the contact interface, including pressure redistribution, tangential force variation, friction-induced vibration, acoustic emission, and even thermal changes. Each sensing modality captures a partial view through its own distinct transduction mechanisms. As a result, modalities provide complementary cues but distinct failure modes, making multimodal sensing a promising direction for robust slip detection. Despite decades of progress since early accelerometer-based artificial skin systems in the 1980s [1], two key challenges still remain unresolved today [2]. ***(i) Precise detection latency*** remains poorly characterized. Detection latency measures the interval between slip onset and the first correct detection, capturing the time needed to accumulate sufficient evidence. This delay often dominates system response time beyond inference speed, yet is rarely reported. Latency also trades off inherently against false alarms. Increasing sensitivity reduces detection delay but raises false positives, and prior work rarely evaluates this trade-off. ***(ii) Generalization*** across unseen objects, sensor units, and robotic platforms remains limited. Most systems are evaluated under conditions closely matching their development setup, leaving it unclear how well they generalize to novel scenarios. This includes shifts in object properties, sensor-level variations, and robotic systems with different contact geometry and actuations.

To address these challenges, we present ***SlipSense***, built on a novel multimodal tactile sensor module (TacV5) integrating a high-density piezoresistive array (Piezo, 240 Hz, 32×32, 155 taxels/cm²) and a high-bandwidth 3-axis MEMS accelerometer (XL, 8 kHz). These two modalities capture complementary physical signatures of contact. Piezo measures a spatial normal pressure distribution and its temporal evolution across the contact surface, while XL captures friction-induced vibrations arising from relative motion at the interface. Crucially, their failure modes are complementary rather than shared: spatial sensing is susceptible to false positives from non-slip pressure changes such as grasp variation, which produce vibration signatures distinct from slip; whereas vibration sensing can be confounded by environmental disturbances that do not introduce quasi-static force redistribution. Jointly processing both streams mitigates each modality's dominant failure mode. Modalities are synchronized across all fingertip sensors, encoded by dedicated backbones, and fused through a causal temporal classifier that performs ternary-class prediction (no-contact, no-slip, slip) at 240 Hz.

Our contributions are: ***(1) A multimodal learning framework*** demonstrating that Piezo and XL signals are complementary and jointly crucial, with fusion achieving 96.7% Macro F1 vs. 81–84% for either modality alone, while reducing false-positive rate below 1.6%. ***(2) A fast slip detection system*** operating at 240 Hz, where 76% of slip events are detected within 23.1 ms, combining detection latency and model inference. ***(3) Cross-domain generalization.*** We conduct experiments on *1.4M* labeled frames covering *37* diverse objects, demonstrating generalization to unseen objects, distinct sensor instances, and zero-shot transfer from a UMI parallel-jaw gripper to a Tesollo dexterous hand.

The remainder of the paper is organized as follows. Section 2 reviews related work across tactile sensing modalities. Section 3 describes the sensor hardware and data collection setup. Section 4 presents the multimodal fusion framework. Section 5 reports experimental results demonstrating high-accuracy, low-latency slip detection and generalization across scenarios.

2 Related Work

Tactile slip detection relies on diverse sensing principles, each with characteristic strengths and failure modes. Optical sensors such as GelSlim [3], GelSight Mini [4], and GelStereo [5] infer slip from displacement fields at high spatial resolution, achieving upto 95% accuracy but at 20–60 Hz, with degradation on unseen objects [4] and smooth surfaces [3, 5]. Among electrical sensors, shear force plays a central role to slip detection through the friction ratio. Capacitive and piezoelectric sensors can directly capture shear related signals, though the latter is limited to dynamic transients [6]. Piezoresistive arrays measure only normal force and often at low spatial resolution. These constraints weaken spatial contact fidelity, leading most prior work to collapse 2D spatial structure into 1D temporal signals for frequency-domain analysis [7, 8]. Our TacV5 instead, achieves a 0.7×0.92 mm pitch ($38.8\times$ higher spatial density than [7] and $6.2\times$ than FlexiTac [9]), enabling joint modeling of spatial pressure structure and its temporal evolution without



Figure 2: *Left:* TacV5 multimodal sensors on UMI grippers and the Tesollo DG-5F dexterous hand at the thumb, index, and middle fingers. *Right:* Data collection platform. The UMI is mounted on a separate table to physically isolate vibrations from the Mark10 and constrained by a fixture to suppress slip-induced shaking. Objects are secured with an adjustable clamp, while controlled slip is generated through programmable vertical crosshead motion. The operator controls only gripper opening and closing. No external camera is required.

explicit shear measurement. Vibration-based methods capture friction-induced oscillations [10] at high bandwidth but respond to environmental disturbances as well [11]. These complementary strengths and weaknesses naturally motivate multimodal fusion. Several multimodal systems have been proposed [12, 13, 14, 15, 16, 17], demonstrating benefits from combining modalities. However, many do not specifically target slip detection, and none evaluates the two critical factors emphasized here, detection latency and generalization.

Detection latency is the operationally relevant metric for closed-loop control, yet is rarely reported especially alongside false-positive rates. Among the few that measure it: BioTac benchmarks against an object-mounted IMU [13], Romeo et al. [18] report 76.7% of trials below 30 ms, Ayral et al. [19] report 20.4 ms across 20 trials on one single object, and Massalim et al. [20] report 17 ms but validate on only three objects. Others report model execution time rather than detection latency [21]. Generalization is equally limited. Accuracy degrades on unseen objects [4], scaling to multifingered hands drops performance [22]. Recent efforts advance generalization over grasp poses on a dexterous hand [21] and across gripper types [11], but do not characterize detection latency alongside transfer performance. To the best of our knowledge, we are the first to jointly evaluate both.

3 System and Data Acquisition

3.1 Sensor Specification

We developed TacV5, a multimodal tactile fingertip module. The primary modality is a 32×32 piezoresistive array (Piezo) with 942 active taxels (0.45×0.45 mm elements at 0.7×0.92 mm pitch, 155 taxels/cm^2) on a flexible polyimide substrate, covering a 0–5 MPa normal pressure range (i.e., 500 N on 1×1 cm^2 area) at 12 bit resolution (i.e., 4096 levels) and 240 Hz. The sensor measures normal pressure only; shear-related dynamics are inferred implicitly by the learning from temporal sequences of pressure maps. The module also includes a 3-axis MEMS accelerometer (XL) sampling at 8 kHz, capturing high-frequency vibration from frictional dynamics at the contact interface. Additional modalities (MEMS microphone, bone-conduction microphone) are included in the module but not used in this work. Lastly, the module communicates over CAN-FD and streams synchronized multimodal data at 240 Hz; we refer to each synchronized sample as a *frame*.

Compared to optical tactile sensors such as GelSight [23] and DIGIT [24], TacV5 offer a thinner form factor, greater mechanical robustness, and requires neither internal illumination nor cameras. Its compact design fits within a standard fingertip pad and remains mechanically independent of gripper actuation, enabling deployment across platforms without sensor-side modification. As shown in Fig. 2, we integrate TacV5 on two platforms: a UMI [25] parallel-jaw gripper with two sensor units (one per jaw), and a Tesollo DG-5F dexterous hand [26] with sensor units to three fingers.

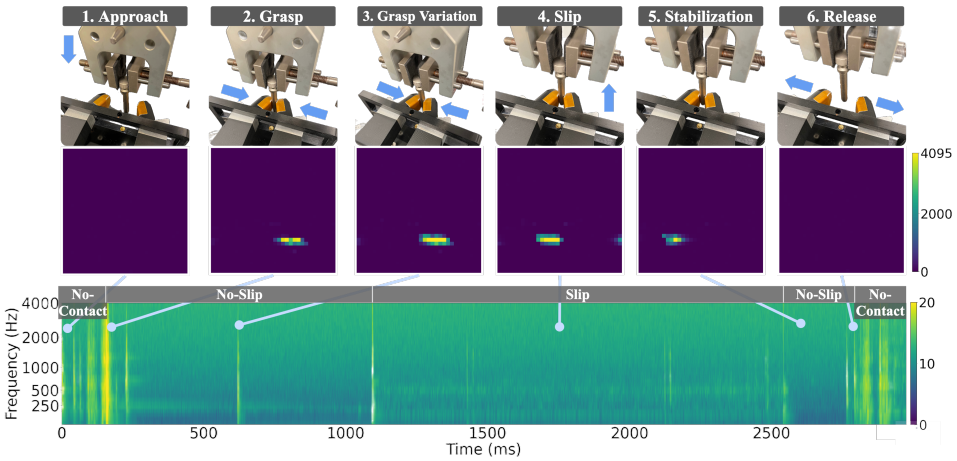


Figure 3: Data collection protocol. Each collection follows a six-step sequence, illustrated with UMI and one sensor's readings: (1) crosshead descends and holds; (2) gripper closes; (3) operator varies grasp force to diversify contact conditions; (4) crosshead ascends to induce slip; (5) crosshead stops; (6) gripper opens. Slip onset and ongoing slip produce distinct XL signatures, while non-slip events such as grasp variation (step 3) and gripper actuation (step 2&6) also generate strong vibration responses. Disambiguating these cases requires spatial pressure cues from Piezo, motivating multimodal fusion.

3.2 Data Collection and Labeling

A fundamental challenge in learning-based slip detection is obtaining reliable ground-truth labels. Existing methods infer slip onset from optical marker displacement [3, 27], friction based criteria on the sensor signal itself [18, 28], or human annotations [5, 21]. Since labels are often derived from the sensor outputs or subjective observation, slip onset remains ambiguous near the transition boundary. Following the idea of directly measuring object displacement with linear encoders [20], we adopt a Mark-10 F105-EM motorized test stand. As shown in Fig. 2, the crosshead clamps one end of the object while the sensor-equipped gripper holds the other. For non-stretchable objects, crosshead displacement directly translates to object slip at the gripper surface. Crosshead position is recorded at 0.02 mm resolution, enabling sub-millimeter slip labels independent of the sensor signal. The non-stretchable constraint applies only during data collection and does not affect deployment.

Each collection follows a fixed six-step sequence (Fig. 3). Slip occurs only during the crosshead ascent (step 4). Slip onset is defined when cumulative crosshead displacement exceeds 0.07 mm and slip termination determined using the same criteria when crosshead stops. All frames within this interval are labeled as *Slip*. Frames with contact but outside this interval are labeled as *No slip*, while frames with all Piezo readings below an object-specific threshold are labeled as *No contact*. During slip, Piezo exhibits spatial pressure shift while XL shows sustained vibration patterns in the log-mel spectrogram. These patterns indicate that discriminative features are already present in the raw sensor streams before any learned encoding.

4 Multimodal Learning

The two sensing modalities, Piezo and XL, differ substantially in temporal granularity, spatial structure, and signal statistics, making naive fusion ineffective. In this section, we present a causal, learning-based SlipSense framework (Fig. 4).

4.1 Problem Formulation

Let S denote the number of sensor units and T the observation window length in frames. At each frame t, sensor $s \in \{1, \ldots, S\}$ produces a Piezo tactile image $\mathbf{p}_{t,s} \in \mathbb{R}^{H \times W}$ and an XL sequence

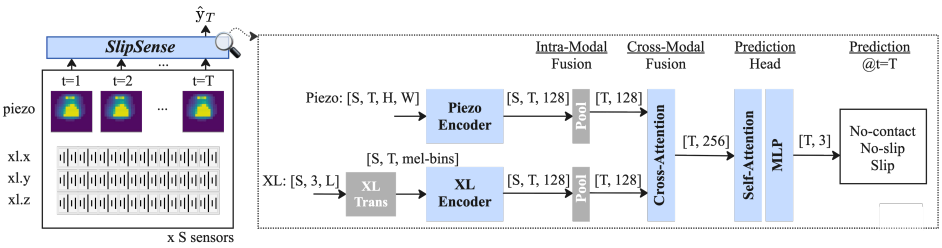


Figure 4: Overview of SlipSense. Given an observation window of length T, Piezo and XL signals are independently encoded into modality specific embeddings at each frame, followed by intra-modal fusion and cross-modal fusion. A causal attention module with a prediction head outputs a 3-class prediction for the last frame at 240 Hz.

$\mathbf{a}_{t,s} \in \mathbb{R}^{3\times L}$, acquired over the time interval $[t-1, t]$. The observation at frame t is defined as $\mathcal{O}_t = \{\, \mathbf{p}_{s,t},\ \mathbf{a}_{s,t} \mid s \in [S] \,\}$.

Each frame carries a label $y_t \in \mathcal{Y} = \{0, 1, 2\}$, corresponding to the *no-contact*, *no-slip*, and *slip* states defined in Sec. 3.2. The objective is to learn a classifier f_θ that takes an observation sequence as input and predicts the label of its last frame T alone:

$$\hat{y}_T = f_\theta(\mathcal{O}_{1:T}), \qquad \hat{y}_T \in \mathcal{Y}. \tag{1}$$

4.2 SlipSense Framework

Piezo encoder. We adopt a lightweight convolutional architecture. The network consists of three sequential blocks, each with a 3×3 convolutions, ReLU activations, and 2×2 max pooling, followed by a fully connected layer that produces a Piezo embedding $\mathbf{z}^p \in \mathbb{R}^d$ with d=128. We also explored larger and pretrained alternatives (Appendix D), but found this compact design both more effective and better suited for low-latency inference.

Accelerometer encoder. Accelerometer signals share the frequency-rich temporal structure of audio, motivating a spectral encoding approach. We high-pass filter at 10 Hz to remove gross gripper motion, then compute per-axis log-mel spectrograms from a causal 16 ms analysis window aligned to each Piezo frame. The three per-axis spectrograms are aggregated via log-sum-exp to obtain a slip-direction-invariant spectrogram (proof in Appendix G). The merged spectrogram exhibits clear slip signatures visible in both low-frequency and high-frequency bands (Fig. 3). We encode it using SSAST-Tiny [29], producing a per-frame embedding $\mathbf{z}^{xl} \in \mathbb{R}^{128}$. Transformation and architectural details are in Appendix B.

Intra and cross modality fusion. Platforms may carry different numbers of sensor units S. To handle this variable-cardinality setting, we aggregate per-sensor embeddings within each modality using element-wise max pooling, i.e., $\mathbf{z}^p_t = \max_{s\in[S]} \mathbf{z}^p_{t,s}$. This produces an embedding independent of sensor count, providing the potential for zero-shot transfer across different platform configurations. To integrate the Piezo and XL features, we apply multi-head cross-attention (MHAttn):

$$\mathbf{z}^{p\prime}_t = \mathrm{LayerNorm}\big(\mathbf{z}^p_t + \mathrm{MHAttn}(Q{=}\mathbf{z}^p_t,\ K{=}\mathbf{z}^{xl}_{\le t},\ V{=}\mathbf{z}^{xl}_{\le t})\big)\,, \tag{2}$$

$$\mathbf{z}^{xl\prime}_t = \mathrm{LayerNorm}\Big(\mathbf{z}^{xl}_t + \mathrm{MHAttn}(Q{=}\mathbf{z}^{xl}_t,\ K{=}\mathbf{z}^{p}_{\le t},\ V{=}\mathbf{z}^{p}_{\le t})\Big)\,, \tag{3}$$

The fused representation concatenates both updated embeddings, $\mathbf{z}^{\mathrm{fused}}_t = \big[\mathbf{z}^{p\prime}_t;\ \mathbf{z}^{xl\prime}_t\big] \in \mathbb{R}^{d_p+d_{xl}}$.

Temporal frame-based predictor. The fused embeddings $\{\mathbf{z}^{\mathrm{fused}}_t\}_{t=1}^T$ are processed by a causal self-attention layer that accumulates temporal context. A prediction head produces per-frame logits $\hat{y}_t \in \mathcal{Y}$. Training supervises only the last frame T, which has access to the full observation window:

$$\mathcal{L} = \mathrm{CE}(\hat{y}_T,\ y_T). \tag{4}$$

During inference, the model can output prediction per incoming frame, enabling real-time operation at 240 Hz when inference completes within one frame interval.

Split		Platform	# Objects	Contact Mode	Sensor Units	# Frames
Train		UMI	28	Fingertip	-	847K
Test	UMI ID	UMI	28	Fingertip	Same	361K
	UMI OOD	UMI	9 (Held-out)	Fingertip	Same	72K
	UMI Fingerpalm	UMI	9 (Held-out)	Fingerpalm	Same	54K
	Tesollo 2-Finger Pinch	Tesollo	9 (Held-out)	2F pinch	Unseen	53K
	Tesollo 3-Finger Pinch	Tesollo	9 (Held-out)	3F pinch	Unseen	45K

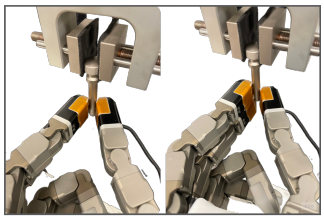


Figure 5: *Left:* Dataset overview. Five test splits progressively evaluate generalization across held-out objects, unseen contact region, and new robotic platforms with physically distinct sensor units, evaluated using Tesollo 2-finger and 3-finger pinch *(Right)*.

Table 1: Multimodal Fusion. Latency is reported as the mean of per-seed median detection latency (in frames) across 3 random seeds; baselines are analytic methods. Fusing Piezo and XL achieves the lowest latency and FPR while substantially improving Macro F1 over either modality alone.

Method	Modality		UMI ID				UMI OOD			
	Piezo	XL	Lat (fr) ↓	FPR (%) ↓	Acc (%) ↑	F1 (%) ↑	Lat (fr) ↓	FPR (%) ↓	Acc (%) ↑	F1 (%) ↑
Baselines	✓		**1.00**	21.69	72.87	74.51	2.00	21.08	73.84	75.29
		✓	**1.00**	22.98	79.41	81.79	2.00	19.31	82.99	84.55
	✓	✓	**1.00**	9.28	81.16	79.96	2.00	7.40	83.33	82.18
SlipSense	✓		20.50	7.54	84.63	83.54	36.17	8.43	82.69	81.91
		✓	4.67	2.14	84.21	81.61	3.00	3.24	84.74	81.33
	✓	✓	2.00	**1.33**	**96.98**	**96.77**	**1.50**	**1.57**	**95.99**	**95.75**

5 Experiments

We evaluate our framework from four perspectives: (1) benefits of multimodal fusion, (2) generalizability across unseen objects, contact regions, sensor units, and robotic platforms, (3) detection latency characterization, and (4) ablations on key design factors affecting performance.

5.1 Experimental Setup

Datasets. We collected a dataset of over 1.4 M frames from 37 objects spanning diverse shapes and surface materials (Appendix Fig. 9), with 28 objects used for training and 9 held out for evaluation. Data were collected under five slip speeds (2–18 mm/s) and varied grasp forces (Appendix Fig. 10). To reduce artifacts and false positives, we additionally include control scenarios without induced slip (Appendix C.1). The resulting dataset contains 57.6% *no-contact*, 23.7% *slip*, and 18.8% *no-slip* frames. We evaluate using five test splits with progressively increasing distribution shift (Fig. 5).

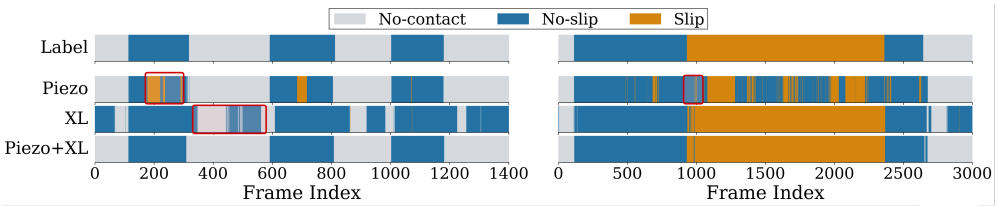
Implementation details. We use a default observation window of T=60 (250 ms). Piezo frames use raw voltage outputs normalized by the 12 bit ADC range of 4096. Modality specific augmentation is applied during training. All experiments are repeated with 3 random seeds and reported as the mean. Full implementation details are provided in Appendix C.2.

Evaluation metrics. We report: *Detection latency*, the median delay from slip onset to first detected slip frame; *False Positive Rate (FPR)*, the fraction of frames incorrectly predicted as slip; *Accuracy*, the fraction of correctly classified frames; and *Macro F1*, the equally weighted class averaged F1.

Baselines. We compare against three analytic baselines: *Piezo only*, adapted from pressure entropy methods [4]; *XL only*, using MFCC features and QDA [30]; and *Piezo-XL*, combining both via AND logic. Full implementation details are provided in Appendix C.3.

5.2 Analysis and Discussion

Piezo and XL are crucially complementary. Table 1 shows an inherent sensitivity-specificity tradeoff in single-modality methods: analytic baselines achieve low latency but high FPR (19–22%), while learned models reduce FPR at the cost of increased latency. Multimodal fusion resolves



(a) Example 1: UMI grasp variation on 3 no-slip contacts (b) Example 2: slip event

Figure 6: Prediction examples. Failure modes are highlighted in red: Piezo can confuse grasp variation with slip and delay slip onset detection, while XL is sensitive to gripper aftershocks during no contact. Fusion suppresses both failure modes, achieving low latency and low false positives.

Table 2: Zero-shot generalization. One single model trained on UMI train set, evaluated zero-shot across three transfer axes: contact location, platform, and finger count.

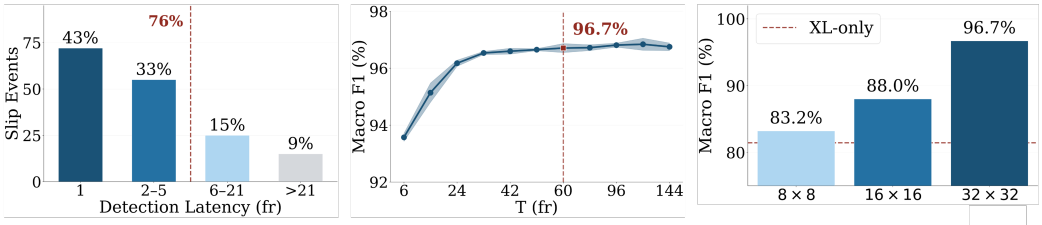
Transfer Scenario	Setup	Lat (fr) ↓	FPR (%) ↓	Acc (%) ↑	F1 (%) ↑
Contact location	UMI Finger Palm	3.00	1.99	94.83	94.56
Platform	Tesollo 2-Finger	3.00	1.14	95.89	94.39
Finger count	Tesollo 3-Finger	5.00	3.12	90.85	87.24

this trade-off, achieving higher Macro F1, lower FPR, and sub-frame latency. Simple AND fusion reduces FPR but over suppresses detections, whereas learned fusion leverages complementary cues for simultaneous gains across all metrics. Fig. 6 illustrates where each modality succeeds and fails. Piezo confuses grasp variation with slip and misses subtle early slip, while XL lacks contact context and responds to environmental vibrations. Fusion combines stable contact grounding from Piezo with the slip onset sensitivity of XL.

Generalization. Unseen-object results (UMI OOD) in Table 1 show additional latency needed for baseline to maintain similar detection performance, indicating slightly more challenging test cases. Notably, our models show minor degradation, suggesting generalizability over object contact patterns. Table 2 further evaluates held-out objects under progressively harder conditions across contact regions, sensor units, and robotic platforms with multi-finger interactions. Fingertip-to-fingerpalm transfer maintains strong performance, indicating the learned features are not tied to a specific contact location. Tesollo presents a more challenging transfer setting with different platform dynamics and physically distinct sensor units. Despite this, the 2-finger pinch achieves 94.39% Macro F1. The 3-finger pinch further alters contact geometry, widening the distribution gap from training, yet performance remains reasonable at 87.24% Macro F1. Across all settings, the model generalizes without targeted fine-tuning, suggesting it captures slip signatures intrinsic to the contact interface rather than artifacts of specific platforms or sensors.

Detection latency characterization. Figure 7(a) shows that slip can be detected almost immediately after onset: 43% of trials are detected at the onset frame and 76% within 20.8 ms. Most long latency cases occur in the 2 mm/s setting. This is the steady crosshead speed, while motion near slip onset is even slower. Even at 2 mm/s, the object moves only 0.5 mm during the 250 ms observation window, close to the tactile array resolution limit, weakening both pressure migration and friction induced vibration. Consistent with this, detection latency decreases with slip speed (Appendix E); at 18 mm/s, over 90% of trials can be detected within 20.8 ms. Model inference averages 2.3 ms on an NVIDIA RTX A4500 (measured over 1,000 runs), well within the 4.17 ms frame budget at 240 Hz. Combined with detection latency, 76% of slip events are detected within 23.1 ms overall.

Choice of observation window T. Fig. 7(b) shows Macro F1 as a function of the observation window T. Performance rises steeply from T=6 fr (93.7%) to T=60 fr (96.7%) and plateaus thereafter. Longer windows also increase variance across seeds, suggesting that excessive temporal context can dilute the local slip cues near the prediction frame. The saturation is physically grounded. At the slowest speed of 2 mm/s, one taxel of displacement requires approximately 450 ms, explaining the modest gains up to T=120 fr (500 ms). At higher slip speeds, T=60 fr (250 ms) already captures sufficient contact evolution, making longer windows largely redundant.



(a) Detection latency characterization (b) Observation window length (c) Piezo resolution

Figure 7: Results on UMI test splits. (a) 76% slip events can be detected within 5 frames (20.8 ms detection latency; 23.1 ms including model inference). (b) Performance saturates at $T{=}60$ fr (250 ms). (c) Fusion performance consistently improves with Piezo resolution.

Piezo spatial resolution benefits. Fig. 7(c) shows the benefit of higher Piezo resolution under multimodal fusion. Lower resolutions ($16{\times}16$ and $8{\times}8$) are generated from native $32{\times}32$ measurements through spatial striding, while keeping the fusion pipeline unchanged. As expected, even an $8{\times}8$ array improves over XL-only, indicating that coarse spatial pressure already provides complementary information. Performance continues to improve with resolution, highlighting the value of fine-grained Piezo sensing and motivating the design towards higher resolution tactile arrays.

5.3 Real-World Deployment and Slip Prevention Strategy

To validate practical utility, we integrate SlipSense into a closed-loop re-grasp controller on the Tesollo hand, where an operator pulls a grasped cable to induce slip. Upon slip detected, the fingers move inward to increase force until the object is secured. Although training data contains only vertical linear slip, SlipSense generalizes to varied slip directions and rotational slip, detecting all slip events in 100/100 trials without false alarms (Appendix F). These detections enable the controller to prevent object loss in 95/100 trials, with a detection-to-peak-force latency of $\sim$50 ms. Importantly, 5 failures stem from the prevention strategy, not missed detections: when the cable is initially grasped near the sensor edge, the coarse inward motion pushes it out despite correct slip detection.

6 Conclusions

We presented SlipSense, a multimodal tactile slip detection framework that fuses $32{\times}32$ piezoresistive arrays with 8 kHz accelerometers. Piezo and XL are complementary: Piezo grounds contact state and suppresses false alarms from environmental vibration, while XL captures friction transients during slip. Their fusion achieves 96.7% Macro F1 with 76% of slip events detected within 23.1 ms. The learned model transfers zero-shot from a UMI parallel-jaw gripper to a Tesollo dexterous hand across unseen objects, distinct sensor units, and different contact configurations, suggesting that the model captures slip signatures intrinsic to the contact interface rather than platform specific artifacts. Integrated into a closed-loop re-grasp controller as a proof of concept, TacV5 and SlipSense demonstrate the potential for fast tactile feedback in dexterous manipulation. Together, these results suggest a promising direction towards native integration of multimodal tactile in dexterous robots.

7 Limitations

While SlipSense demonstrates low-latency detection and zero-shot cross-platform transfer, several limitations remain. Data are collected under a single linear slip axis; rotational and multi-directional slip are successfully detected in real-world deployment (100/100 trials) but without quantitative latency characterization, and in-motion slip detection during active robot manipulation remains to be quantitatively validated as well. The piezoresistive array measures only normal pressure and infers shear implicitly from temporal frame sequences; direct shear sensing may provide further discriminative signal. Although evaluated on 37 objects, broader object and environmental coverage are natural next steps. Finally, inference is benchmarked on a GPU workstation; deployment on embedded hardware would enable a fully self-contained slip detection module.

References

[1] R. Howe and M. Cutkosky. Sensing skin acceleration for slip and texture perception. In *Proceedings, 1989 International Conference on Robotics and Automation*, pages 145–150 vol.1, 1989. doi:10.1109/ROBOT.1989.99981.

[2] X. Zhang, L. Wu, S. He, Z. Sha, Y. Huang, S. Wu, D. Chu, C. H. Wang, and S. Peng. Recent advances of slip sensors for smart robotics. *Advanced Materials Technologies*, page e02251, 2026. doi:10.1002/admt.202502251.

[3] S. Dong, D. Ma, E. Donlon, and A. Rodriguez. Maintaining grasps within slipping bound by monitoring incipient slip. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 3818–3824, 2019. arXiv:1810.13381.

[4] X. Hu, A. Venkatesh, Y. Wan, G. Zheng, N. Jawale, N. Kaur, X. Chen, and P. Birkmeyer. Learning to detect slip through tactile estimation of the contact force field and its entropy properties. *Mechatronics*, 104:103258, 2024.

[5] S. Cui, S. Wang, R. Wang, S. Zhang, and C. Zhang. Learning-based slip detection for dexterous manipulation using gelstereo sensing. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13691–13700, 2023.

[6] R. A. Romeo and L. Zollo. Methods and sensors for slip detection in robotics: A survey. *IEEE Access*, 8:73027–73050, 2020. doi:10.1109/ACCESS.2020.2987849.

[7] C. Schürmann, M. Schöpfer, R. Haschke, and H. Ritter. A high-speed tactile sensor for slip detection. In *Towards Service Robots for Everyday Environments*, volume 76 of *Springer Tracts in Advanced Robotics*, pages 403–415. Springer, 2012. doi:10.1007/978-3-642-25116-0_27.

[8] R. A. Romeo, C. M. Oddo, M. C. Carrozza, E. Guglielmelli, and L. Zollo. Slippage detection with piezoresistive tactile sensors. *Sensors*, 17(8):1844, 2017. doi:10.3390/s17081844.

[9] B. Huang, J. Xu, I. Akinola, W. Yang, B. Sundaralingam, R. O'Flaherty, D. Fox, X. Wang, A. Mousavian, Y.-W. Chao, et al. Vt-refine: Learning bimanual assembly with visuo-tactile feedback via simulation fine-tuning. In *Conference on Robot Learning*, pages 2484–2500. PMLR, 2025.

[10] E. G. M. Holweg, H. Hoeve, W. Jongkind, L. Marconi, C. Melchiorri, and C. Bonivento. Slip detection by tactile sensors: Algorithms and experimental results. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 3234–3239, 1996.

[11] M. Cravetz, P. Vyas, C. Grimm, and J. R. Davidson. Slip detection for compliant robotic hands using inertial signals and deep learning. *Frontiers in Robotics and AI*, 12:1698591, 2025. doi:10.3389/frobt.2025.1698591.

[12] J. Reinecke, A. Dietrich, F. Schmidt, and M. Chalon. Experimental comparison of slip detection strategies by tactile sensing with the BioTac on the DLR hand arm system. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 2742–2748, 2014.

[13] Z. Su, K. Hausman, Y. Chebotar, A. Molchanov, G. E. Loeb, G. S. Sukhatme, and S. Schaal. Force estimation and slip detection/classification for grip control using a biomimetic tactile sensor. In *Proc. IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 297–303, 2015.

[14] Z. Xu, Z. Si, K. Zhang, O. Kroemer, and Z. Temel. A multi-modal tactile fingertip design for robotic hands to enhance dexterous manipulation. *arXiv preprint arXiv:2510.05382*, 2025.

[15] N. Komeno and T. Matsubara. Incipient slip detection by vibration injection into soft sensor. *IEEE Robotics and Automation Letters*, 9(4):3251–3258, 2024.

[16] D. R. Shepherd, P. Husbands, A. Philippides, and C. Johnson. Texture and friction classification: Optical TacTip vs. vibrational piezoeletric and accelerometer tactile sensors. *Sensors*, 25 (16):4971, 2025. doi:10.3390/s25164971.

[17] C. Yu, M. R. Cavallari, and I. Kymissis. Tri-modal thin-film flexible electronic skin to augment robotic grasping. In *2018 IEEE Micro Electro Mechanical Systems (MEMS)*, pages 886–888, 2018. doi:10.1109/MEMSYS.2018.8346698.

[18] R. A. Romeo, C. Lauretti, C. Gentile, E. Guglielmelli, and L. Zollo. Method for automatic slippage detection with tactile sensors embedded in prosthetic hands. *IEEE Transactions on Medical Robotics and Bionics*, 3(2):485–497, 2021. doi:10.1109/TMRB.2021.3060032.

[19] T. Ayral, S. Aloui, and M. Grossard. Reactive slip control in multifingered grasping: Hybrid tactile sensing and internal-force optimization. *arXiv preprint arXiv:2602.16127*, 2026.

[20] Y. Massalim, Z. Kappassov, and H. A. Varol. Deep vibro-tactile perception for simultaneous texture identification, slip detection, and speed estimation. *Sensors*, 20(15):4121, 2020. doi:10.3390/s20154121.

[21] C. Zhao, Y. Yu, Z. Ye, Z. Tian, Y. Zhang, and L.-L. Zeng. Universal slip detection of robotic hand with tactile sensing. *Frontiers in Neurorobotics*, 19:1478758, 2025. ISSN 1662-5218. doi:10.3389/fnbot.2025.1478758.

[22] J. W. James and N. F. Lepora. Slip detection for grasp stabilization with a multifingered tactile robot hand. *IEEE Transactions on Robotics*, 37(2):506–519, 2021. doi:10.1109/TRO.2020.3031245.

[23] W. Yuan, S. Dong, and E. H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[24] 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*, 5(3):3838–3845, 2020.

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

[26] TESOLLO Inc. Dg-5f-m humanoid robotic hand. `https://en.tesollo.com/dg-5f-m/`, 2025. Accessed: 2026-05-14.

[27] J. W. James, N. Pestell, and N. F. Lepora. Slip detection with a biomimetic tactile sensor. *IEEE Robotics and Automation Letters*, 3(4):3340–3346, 2018. doi:10.1109/LRA.2018.2852797.

[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*, pages 885–915. PMLR, 2025.

[29] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10699–10709, 2022.

[30] Y. Yoo, C.-Y. Lee, and B.-T. Zhang. Multimodal anomaly detection based on deep auto-encoder for object slip perception of mobile manipulation robots. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, page 11443–11449. IEEE, May 2021. doi:10.1109/icra48506.2021.9561586. URL `http://dx.doi.org/10.1109/ICRA48506.2021.9561586`.

[31] U. Yoo, Y. Mao, J. Oh, and J. Ichnowski. A-slip: Acoustic sensing for continuous in-hand slip estimation. *arXiv preprint arXiv:2604.08528*, 2026.

[32] Y. Gong, Y.-A. Chung, and J. Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. doi:10.21437/Interspeech.2021-698.

[33] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[34] 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*, 38: 153783–153812, 2026.

[35] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Appendix

This appendix provides supplementary material in seven parts: customized UMI setup (App. A), XL transformation and encoder details (App. B), experimental setup (App. C), ablations on encoder architectures and inference cost (App. D), detailed detection latency characterization by slip speed (App. E), real-world SlipSense deployment on UMI and Tesollo (App. F), and proof of direction invariance of merged log-mel spectrogram (App. G).

A Customized UMI

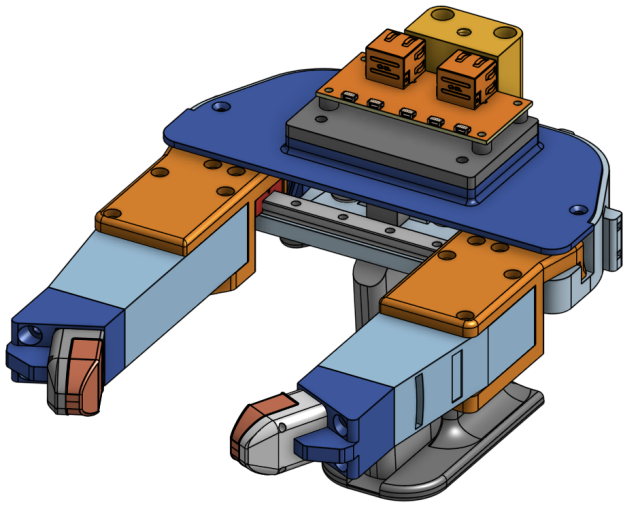


Figure 8: A customized UMI was developed to augment the data-collection interface with tactile sensing. Two redesigned finger-adapter assemblies were mounted to the UMI body, enabling attachment of dexterous-hand fingertips equipped with tactile sensors, shown in orange. A top-mounted aggregation board, also shown in orange, was installed at the center of the main body to consolidate and synchronize data streams from the two tactile fingers.

B Accelerometer Transformation and Encoder Details

For each axis $k \in \{x, y, z\}$, we compute a causal log-mel spectral frame $\mathbf{M}_{t,k} \in \mathbb{R}^F$ from a 16 ms analysis window ending at frame t, with a frame shift of $1/240$ s aligned one-to-one with each Piezo frame. Each spectral frame uses only past and current observations without future information. Contact vibrations project differently onto each accelerometer axis depending on slip direction, making per-axis spectrograms direction-dependent. While this property can be useful for slip direction estimation [31], our data are collected under a fixed direction and we aim for direction-invariant generalization. We therefore aggregate across axes using log-sum-exp:

$$\mathbf{M}_t = \log \sum_{k \in \{x,y,z\}} \exp(\mathbf{M}_{t,k}) \in \mathbb{R}^F, \tag{5}$$

This yields a provably direction-invariant representation (Appendix G).

The merged spectrogram is encoded by SSAST-Tiny [29], an AST-family model [32]. Contact signals exhibit sparse spectral patterns compared to speech, so we reduce the standard 128 mel bins to F=16, using non-overlapping 16×1 patches that collapse the frequency dimension to a

single embedding per time step. The patch-embedding projection is reinitialized to accommodate the reduced input resolution, while all transformer layers retain pretrained weights. The encoder outputs a per-frame embedding $\mathbf{z}^{xl} \in \mathbb{R}^{128}$.

C Experimental Details

C.1 Dataset

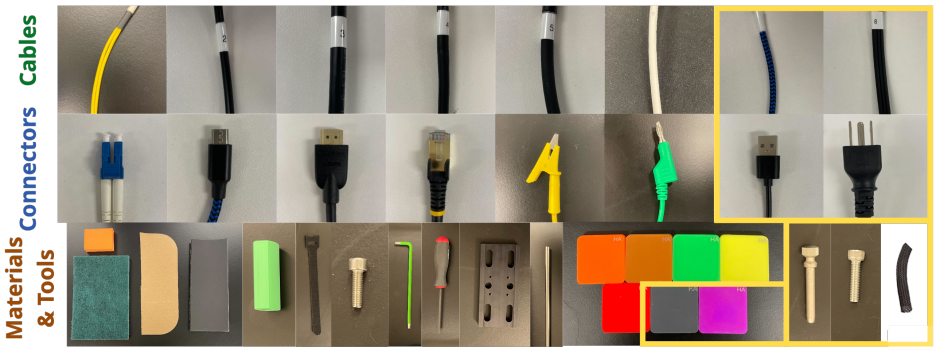


Figure 9: 37 objects spanning cables, connectors, materials, and tools. We train on 28 objects and reserve 9 held-out objects for evaluation (yellow boxes). The held-out set further includes Shore hardness test blocks, with hardness levels 81 HA and 88 HA unseen from training to evaluate generalization across material properties.

Slip data were collected under five crosshead speeds ranging from 2–18 mm/s and a human natural range of grasp forces. Fig. 10 shows the distribution of total contact pressure across the training and test sets. The training set is collected by a human operator who intentionally varies grasp force, naturally reflecting the range of pressures encountered in human-operated grasping. The resulting training distribution (blue) covers the full range of contact forces observed during testing (pink), with both concentrated below 200 kPa·taxels and a long tail extending beyond 400 kPa·taxels. This confirms that the model is not evaluated outside its trained pressure regime.

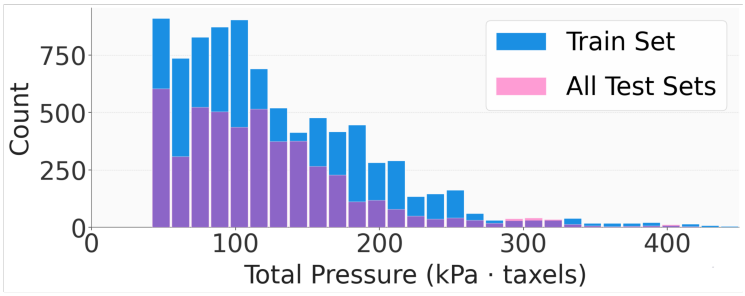


Figure 10: Contact pressure coverage. Distribution of total pressure (kPa·taxels) for the training set (blue) and all test sets (pink). Training data is collected by a human operator intentionally varying grasp force, naturally reflecting the human-operated pressure range and covering the full test distribution.

To reduce Mark10 induced artifacts and suppress false positives, we include three control scenarios in the training set, each with 5 trials:

(1) Crosshead motion with the UMI gripper not in contact with any object;

(2) Crosshead motion while the UMI gripper holds an object not linked to the crosshead;

(3) Free-grasping and pick-and-place motions using the UMI without induced slip.

C.2 Implementation Details

We use a default observation window length of $T{=}60$. Piezo augmentation includes random rotation up to 360°, spatial rolling up to $\pm 30\%$, horizontal and vertical flipping, and additive Gaussian noise with $\sigma = 0.001$. XL log-mel spectrograms are standardized across frequency bins per frame and augmented with frequency masking of up to 5 bins applied with 50% probability, together with additive uniform noise. All models are trained for 10 epochs with learning rate 1×10^{-5} and batch size 32. Training uses a 5 epoch linear warmup followed by MultiStepLR decay with $\gamma = 0.5$. We apply Exponential Moving Average for stable inference.

C.3 Baseline Details

We compare against three analytic baselines: *(1) Piezo-only*, adapting the entropy-based slip indicator from Hu et al. [4] by replacing GelSight marker motion with frame-to-frame pressure differences ΔP_t on the 32×32 grid, restricted to active contact regions exceeding 5% of peak pressure. *(2) XL-only*, treating accelerometer signals as acoustic data and following Yoo et al. [30] to compute 13 MFCC features over 26 mel filterbanks with Quadratic Discriminant Analysis (QDA) for classification. *(3) Piezo-XL*, combining predictions from both baselines via logical AND. All thresholds are fitted on the training set and applied directly to test data.

D Ablation Study on Encoder Architectures

The choice of per-modality encoder directly affects how well each signal stream is represented before fusion. A weak encoder can bottleneck the entire pipeline regardless of fusion quality. Since single-modality performance is substantially lower than the fused model (Table 1), ablating encoders in isolation would conflate encoder quality with the absence of the complementary modality. We therefore ablate in the full Piezo+XL fused setting, keeping the fusion pipeline and one modality's encoder fixed while varying the other. This isolates each encoder's contribution to the fused performance and better reflects the deployment configuration. The encoder candidates for each modality are described below.

Piezo Encoder (ResNet-18) ResNet-18 [33] processes piezo tactile images (32×32) through successive convolutional blocks with skip connections, producing a 128-dimensional feature representation via global average pooling.

Piezo Encoder (ViT-MAE) Tactile images contain rich spatial structure that benefits from self-supervised representation learning [34]. We pretrain a ViT-based encoder on approximately 2 million piezoresistive contact frames: ∼2M simulated grasps (72 objects) from our simulator across 72 objects, and ∼30K real-world frames from operator-held UMI grasping across 10 objects. We use a masked autoencoding (MAE) [35], treating each pressure image $\mathbf{p}_{t,s} \in \mathbb{R}^{H \times W}$ as a single-channel image divided into non-overlapping $4{\times}4$ patches, yielding an $8{\times}8$ grid of 64 tokens. During pretraining, 60–80% of patches are randomly masked in 95% of training samples; the asymmetric encoder (8 transformer layers, 128-dimensional hidden size, 8 attention heads) processes only visible patches, while a lightweight decoder (4 layers, 64-dimensional hidden size, 4 attention heads) reconstructs masked patches from encoded visible tokens and learnable mask tokens. Reconstruction loss (MSE on raw pixel values) is computed only on masked patches, encouraging spatially coherent representations from partial observations. Pretraining data is collected from both real grasps and physics-based simulation across diverse object geometries. After pretraining, masking is disabled and the encoder operates on all 64 patches to extract the `[CLS]` token as the force embedding $\mathbf{z}^p \in \mathbb{R}^d$, where d = 128.

Three variants are considered when training SlipSense: *i) scratch*, training the encoder from random initialization; *ii) freeze*, keeping the pretrained parameters fixed; and *iii) finetune*, initializing from pretrained weights and updating them during training.

Accelerometer Encoder (MLP) A compact two-layer multilayer perceptron for processing low-dimensional accelerometer features. The network transforms input vectors through a hidden layer with 16 units and ReLU activation, followed by dropout regularization, and outputs feature embeddings of configurable dimension. This architecture provides a parameter-efficient baseline for encoding pre-processed accelerometer signals.

Accelerometer Encoder (1D CNN) A causal convolutional network for temporal sequence modeling of accelerometer data. The architecture employs three 1D convolutional layers with batch normalization and ReLU activations: an initial temporal convolution with causal padding (left-padding only) to prevent information leakage from future timesteps, followed by two 1×1 convolutions. The network outputs per-frame feature embeddings, preserving the temporal structure of the input sequence.

Table 3 reports results on UMI OOD. Execution time is reported as the mean over 1000 runs on an NVIDIA RTX A4500 GPU. For Piezo, all ViT-MAE variants outperform ResNet-18, confirming that spatial self-supervised pretraining benefits pressure-map encoding. The frozen encoder slightly outperforms finetuning, suggesting that pretrained representations already capture the relevant spatial structure and that end-to-end adaptation risks overfitting. The Custom CNN matches ViT-MAE variants across all metrics while being substantially lighter, indicating that 32×32 tactile images do not require the capacity of a vision transformer. For XL, the contrast is sharper. An MLP operating on flattened spectrograms collapses to 29-frame median latency, showing that temporal structure in the spectrogram is essential and cannot be captured by a position-agnostic architecture. The 1D CNN recovers most of the performance, but Tiny-SSAST achieves the best results across all metrics, validating the transfer of audio-domain pretrained weights to accelerometer spectrograms.

Table 3: Encoder architecture ablation. Each row varies one modality's encoder while fixing the other and the fusion pipeline (Piezo+XL). Evaluated on UMI OOD. Selected encoders (†) achieve the best overall performance.

Modality	Architecture	Lat (fr) ↓	FPR (%) ↓	Acc (%) ↑	F1 (%) ↑	Execution Time (ms)↓
Piezo	ResNet-18	2.00	4.19	93.04	92.91	3.08
	ViT-MAE (scratch)	2.67	2.26	95.38	95.33	5.31
	ViT-MAE (freeze)	2.67	2.02	95.75	95.51	5.33
	ViT-MAE (finetune)	2.00	3.17	95.07	95.03	5.32
	Custom CNN†	**1.50**	**1.57**	**95.99**	**95.75**	2.33
XL	MLP	30.00	4.04	90.69	90.38	1.07
	1D CNN	2.50	2.30	94.88	94.68	1.33
	Tiny-SSAST†	**1.50**	**1.57**	**95.99**	**95.75**	2.33

E Detection Latency by Slip Speed

Table 4 and Fig. 11 break down detection latency by slip speed. Faster crosshead motion produces larger friction transients in XL and faster pressure redistribution in Piezo, both of which strengthen the slip signature. At 18 mm/s, over 90% of events are detected within 5 frames (20.8 ms). Slow slip at 2 mm/s is the most challenging condition, with a long tail extending to 105 frames, reflecting cases where gradual pressure migration takes time to accumulate sufficient evidence for detection.

F Real-World Deployment

We validate SlipSense in two real-world settings (Fig. 12 and Fig. 13). First, an operator holds a free-standing UMI gripper — as opposed to the table-mounted configuration used during data collection — introducing hand-induced motion that may excite the XL and varies Piezo contact pressure subtly. Despite these disturbances, SlipSense produces no false alarms, owing to the control scenarios in the training set. Second, we evaluate slip detection on the Tesollo hand across varied conditions,

Table 4: Detection Latency by Slip Speed. Median and mean detection delay (frames) across five programmed crosshead speeds on UMI test episodes. n denotes the number of slip events per speed. One frame = 4.17 ms at 240 Hz.

Speed	n	Median (fr) ↓	Mean (fr) ↓	Max (fr) ↓
2 mm/s	36	3.0	14.9	106
5 mm/s	31	3.0	6.8	39
10 mm/s	37	1.0	3.4	27
15 mm/s	32	2.5	5.8	35
18 mm/s	31	1.0	2.9	20

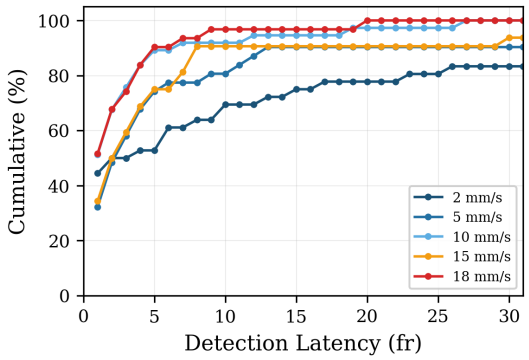


Figure 11: Detection latency CDF by slip speed. Cumulative distribution of slip-onset detection delay (in frames) across five crosshead speeds on UMI test episodes. Faster slip produces stronger friction transients and pressure migration, leading to earlier detection. At 18 mm/s, over 90% of events are detected within 5 frames (20.8 ms).

including different cable types, different pull directions, and rotational slip. SlipSense correctly identifies all slip events in 100 trials.

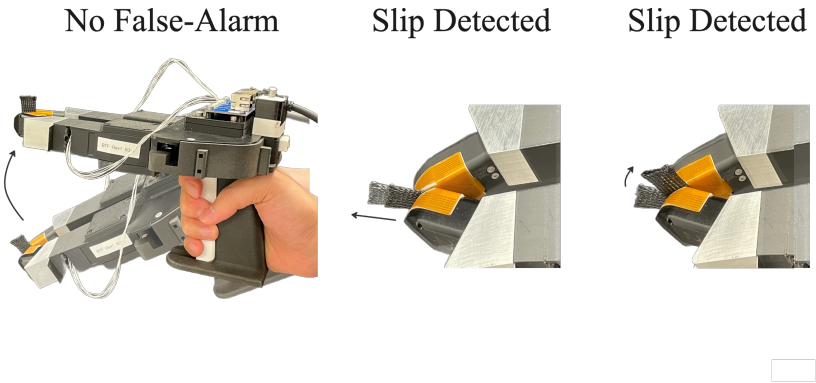


Figure 12: Real-world slip detection on UMI. Free UMI gripper held by an operator. Hand-induced motion may excite the XL and varies Piezo contact subtly, but *SlipSense* produces no false alarms (left). Translational slip (center) and rotational slip (right) can be correctly detected.

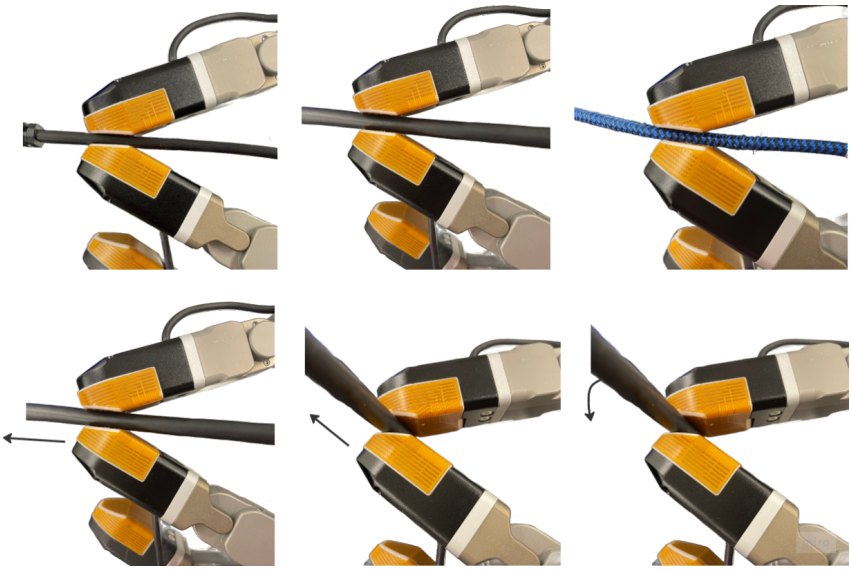


Figure 13: Real-world slip detection on the Tesollo hand. Training data contains only vertical linear slip. Through Piezo data augmentation (Appendix. C.2) and direction-invariant XL encoding (Appendix. G), *SlipSense* generalizes to unseen directions and rotational slip. Top: hand grasp configurations on three representative cables; trials include additional types beyond those shown. Bottom: varied pull directions and rotational slip. *SlipSense* correctly identifies slip in all 100/100 trials.

G Proof of Direction Invariance of Merged Log Mel Spectrogram

Notation. Let $s(t) \in \mathbb{R}$ denote the scalar vibration signal at the contact interface, and let $\mathbf{d} = (d_x, d_y, d_z)^\top \in \mathbb{R}^3$ denote the unit direction vector of slip or motion. The three-axis accelerometer measurements are $x(t)$, $y(t)$, $z(t) \in \mathbb{R}$, with corresponding STFT coefficients $X(f,\tau)$, $Y(f,\tau)$, $Z(f,\tau)$, $S(f,\tau) \in \mathbb{C}$. The k-th mel filterbank weight at frequency f is $m_k(f) \geq 0$, and $\mathcal{M}_k^s(\tau)$ denotes the k-th mel filterbank output applied to $|S(f,\tau)|^2$. The per-axis log-mel filterbank output is written $\mathcal{F}_k^{(i)}(\tau)$ for axis $i \in \{x, y, z\}$.

Assumptions.

1. **(Linear projection.)** The vibration at the contact interface is a scalar signal $s(t)$ propagating along a single fixed direction $\mathbf{d}$, such that the acceleration measured at the sensor satisfies:
$$x(t) = d_x \cdot s(t), \quad y(t) = d_y \cdot s(t), \quad z(t) = d_z \cdot s(t). \tag{6}$$
2. **(Unit direction.)** $\mathbf{d}$ is a unit vector:
$$d_x^2 + d_y^2 + d_z^2 = 1. \tag{7}$$
3. **(Quasi-stationary direction.)** $\mathbf{d}$ is constant within each STFT analysis window.
4. **(Single vibration source.)** The sensor receives vibration from a single dominant source, such that superposition of independent multi-directional sources is negligible.

Proposition 1. *Under assumptions **A1–A4**, the log-mel spectrogram computed from the summed power spectrum of the three accelerometer axes is invariant to the slip or motion direction* $\mathbf{d}$.

Proof. **Step 1: STFT linearity.** By linearity of the STFT and **A1**:
$$X(f,\tau) = d_x \cdot S(f,\tau), \quad Y(f,\tau) = d_y \cdot S(f,\tau), \quad Z(f,\tau) = d_z \cdot S(f,\tau). \tag{8}$$

Step 2: Per-axis power spectra.
$$|X(f,\tau)|^2 = d_x^2\,|S(f,\tau)|^2, \quad |Y(f,\tau)|^2 = d_y^2\,|S(f,\tau)|^2, \quad |Z(f,\tau)|^2 = d_z^2\,|S(f,\tau)|^2. \tag{9}$$

Step 3: Summed power spectrum. Summing across axes and applying **A2**:
$$|X(f,\tau)|^2 + |Y(f,\tau)|^2 + |Z(f,\tau)|^2 = \underbrace{\left(d_x^2 + d_y^2 + d_z^2\right)}_{=1} |S(f,\tau)|^2 = |S(f,\tau)|^2. \tag{10}$$

The direction-dependent coefficients cancel exactly.

Step 4: Mel filterbank. Applying the linear mel filterbank with non-negative weights to the summed power:
$$\mathcal{M}_k(\tau) = \sum_f m_k(f)\Big(|X|^2 + |Y|^2 + |Z|^2\Big) = \sum_f m_k(f)\,|S(f,\tau)|^2 = \mathcal{M}_k^s(\tau). \tag{11}$$

Step 5: Log.
$$\log \mathcal{M}_k(\tau) = \log \mathcal{M}_k^s(\tau), \tag{12}$$
which is identical to the log mel spectrogram of $s(t)$ alone, independent of $\mathbf{d}$. □

Corollary 1 (LogSumExp equivalence)**.** *When per-axis log-mel features are computed separately, the direction-invariant log-mel spectrogram is equivalently obtained via* LogSumExp *across axes. Specifically, for each axis* $i \in \{x, y, z\}$*:*
$$\mathcal{F}_k^{(i)}(\tau) = \log\big(d_i^2 \cdot \mathcal{M}_k^s(\tau)\big) = 2\log|d_i| + \log \mathcal{M}_k^s(\tau). \tag{13}$$

Applying LogSumExp across axes:

$$\log \sum_{i\in\{x,y,z\}} \exp\Big(\mathcal{F}_k^{(i)}(\tau)\Big) = \log \left[\underbrace{\left(d_x^2 + d_y^2 + d_z^2\right)}_{=1} \mathcal{M}_k^s(\tau) \right] = \log \mathcal{M}_k^s(\tau). \quad \square \tag{14}$$

Hence, LogSumExp applied across per-axis outputs is a computationally efficient and numerically stable realization of the direction-invariant log-mel spectrogram.