2025

VibTac: A high-resolution high-bandwidth tactile sensing finger for multi-modal perception in robotic manipulation

Sheeraz Athar

Xinwei Zhang

Jun Ueda

Ye Zhao

Yu She

VibTac: A high-resolution high-bandwidth tactile sensing finger for multi-modal perception in robotic manipulation

Sheeraz Athar[†,1], Xinwei Zhang[†,1], Jun Ueda[2], Ye Zhao[2], Yu She[*,1]

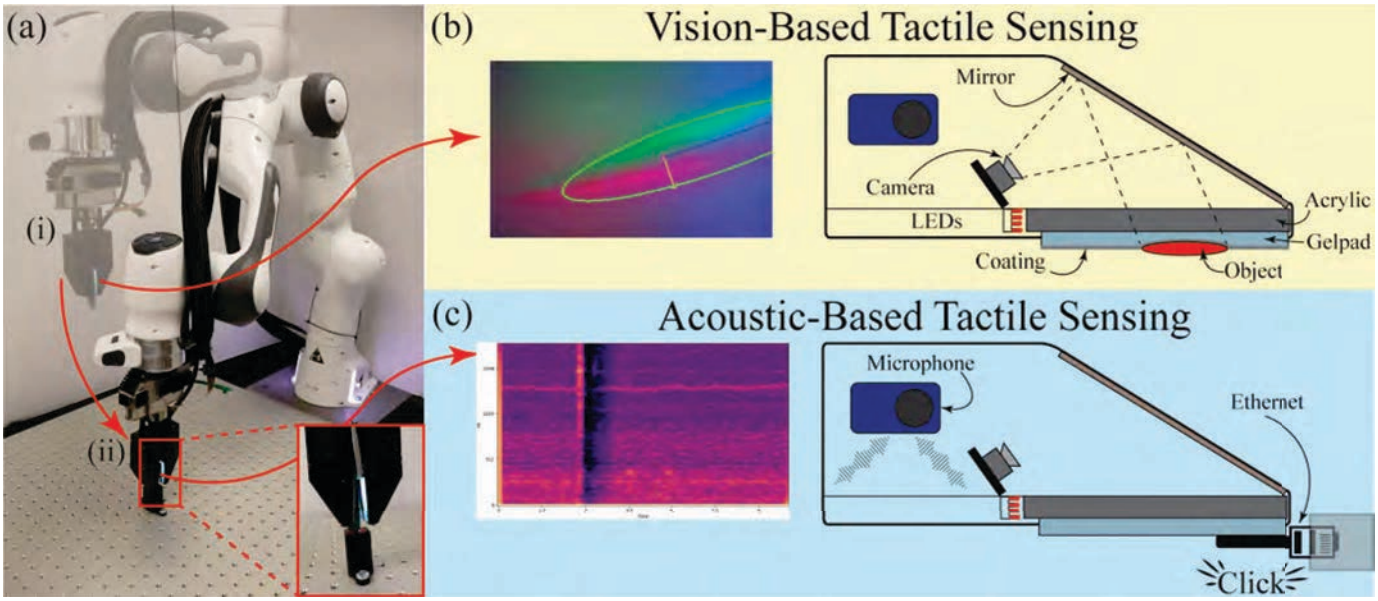

Fig. 1: **Overview of the proposed sensor**: The figure demonstrates the dual-modal sensing of the proposed sensor integrated with a *Franka Panda* robot. (a) In (i), the robot grips an Ethernet cable in a random orientation. Vision-based tactile sensing (shown in (b)) determines the *in-hand* pose and adjusts orientation. The robot then moves to (ii) for insertion. During this phase, acoustic-based tactile sensing (shown in (c)) detects the "click" sound upon full insertion, signaling the robot to stop, confirming task completion.

***Abstract*—Tactile sensing is pivotal for enhancing robot manipulation abilities by providing crucial feedback for localized information. However, existing sensors often lack the necessary resolution and bandwidth required for intricate tasks. To address this gap, we introduce *VibTac*, a novel multi-modal tactile sensing finger designed to offer high-resolution and high-bandwidth tactile sensing simultaneously. *VibTac* seamlessly integrates vision-based and vibration-based tactile sensing modes to achieve high-resolution and high-bandwidth tactile sensing respectively, leveraging a streamlined human-inspired design for versatility in tasks. This paper outlines the key design elements of *VibTac* and its fabrication methods, highlighting the significance of the Elastomer Gel Pad (EGP) in its sensing mechanism. The sensor's multi-modal performance is validated through 3D reconstruction and spectral analysis to discern tactile stimuli effectively. In experimental trials, *VibTac* demonstrates its efficacy by achieving over 90% accuracy in insertion tasks involving objects emitting distinct sounds, such as ethernet connectors. Leveraging vision-based tactile sensing for object localization and employing a deep learning model for "click" sound classification, *VibTac* showcases its robustness in real-world scenarios. Video of the sensor working can be accessed at https://youtu.be/kmKIUlXGroo.**

***Index Terms*—tactile sensing, vision-based tactile, vibration-based tactile, manipulation.**

*Corresponding Author.
†Equal Controbution.
[1]Edwardson School of Industrial Engineering, Purdue University, West Lafayette, IN, USA.
[2]George W. Woodruff School of Mechanical Engineering, Gerogia Institute of Technology, Atlanta, GA, USA.
sathar@purdue.edu, zhan4645@purdue.edu, shey@purdue.edu.
ye.zhou@me.gatech.edu, jun.ueda@me.gatech.edu

I. Introduction

Robots are increasingly becoming integral to both our personal and professional environments. The rapid advancement in robotic technology is expected to lead to an unprecedented demand for their deployment. One of the challenging applications in robotics is complex manipulation, a skill humans routinely employ to complete various tasks [1]. Enabling robots to execute complex hand manipulation resembling that of humans is crucial for their widespread adoption in human spaces.

An essential aspect of successful human manipulation is the ability to sense the environment [2]. Humans utilize various sensing cues to gather pertinent information for task execution. Tactile sensing, in particular, plays a vital role in manipulation

by providing localized information about the object being handled [3]. Similarly, tactile sensing is of paramount importance for robots, furnishing them with the necessary information to manipulate objects effectively [4], [5].

For robotic tactile sensing, one of the biggest hurdles is to provide robots with both high-resolution and high-bandwidth tactile information. Currently, most tactile sensors either provide high-resolution information [6], [7] or high-bandwidth [8]. Equipping robots with both high-resolution and high-bandwidth tactile data is still not trivial for tactile sensors.

On the other hand, human tactile sensing provides both high-resolution and high-bandwidth tactile sensing by combining different mechanisms [9]. Humans also leverage acoustic sensing to infer critical information during manipulation tasks. Acoustic cues help distinguish objects, detect faults, confirm task completion, and assess structural integrity [10]. For example, the distinct "click" sound when inserting an Ethernet cable signals successful assembly. This illustrates how humans integrate tactile and acoustic sensing to manage tasks involving diverse contact dynamics.

Researchers frequently use tactile sensing for robotic manipulation tasks [11]–[13]. However, there has been limited exploration into developing comprehensive methods that equip robots with both high-resolution and high-bandwidth tactile sensing. To achieve human-like dexterity, robots should be equipped with this technology, which they currently lack.

To address this issue, we introduce in this paper a novel tactile sensing finger called *VibTac*. *VibTac* combines vision and acoustic-based tactile sensing to offer high-resolution and high-bandwidth tactile sensing within a single hardware setup. Fig. 1 presents an overview of the proposed sensing finger, which will be used for manipulation in an actual setup.

The finger comprises a vision-based tactile sensing module for high-resolution sensing. Additionally, it incorporates an accelerometer and a microphone to enable high-bandwidth tactile sensing. The sensor captures both micro-slip (small, localized shifts in the contact area) and gross-slip (larger, more noticeable movements) by detecting their unique vibration patterns. In robotic grasping, micro-slip happens when an object, like a pen, slightly shifts within the grip without slipping out completely, while gross slip occurs when the object moves significantly or falls entirely from the hand [14]. These two types of slip produce different vibration frequencies: micro-slip generates high-frequency vibrations (>200 Hz), which are detected by the microphone, while gross slip results in lower-frequency vibrations (<200 Hz), best captured by the accelerometer [15], [16]. *VibTac* utilizes its vision-based tactile sensing capabilities for *in-hand* localization and employs acoustic signals to infer structural integrity, such as the successful insertion of parachute buckles or the completion of tasks like the successful insertion of an Ethernet cable.

The main contributions of the paper are as follows: It showcases significant advancements in robotic tactile sensing and manipulation through the introduction of a multi-modal tactile sensing finger, offering both high-resolution and high-bandwidth tactile feedback, equips robots with refined tactile information, enabling them to execute long-horizon manipulation tasks. Additionally, this paper presents a software pipeline

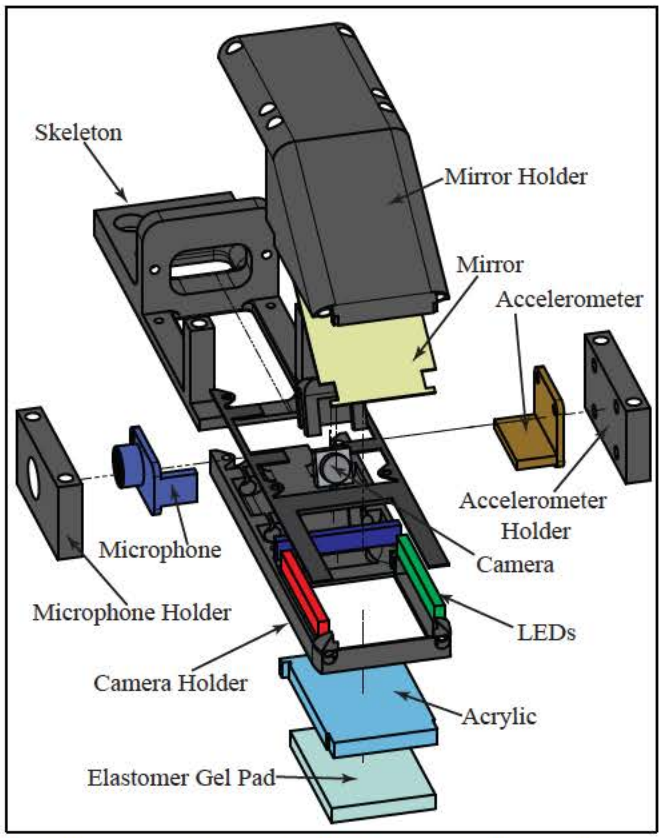

Fig. 2: **Design**: Exploded view of the *VibTac* finger, showing its main components. The camera holder contains the camera, LED acrylic, and elastomer gel pad. The mirror holder secures the mirror, reflecting gel deformations to the camera. The microphone holder houses the microphone, and the accelerometer holder secures the accelerometer. All components are connected by a central skeleton.

that converts raw tactile data into crucial state feedback for tasks such as determining the precise position and orientation of hand-held objects and assessing the success of insertions. The study also emphasizes the robust execution of a class of robotic manipulation tasks, made possible by integrating both high-resolution and high-bandwidth tactile feedback, thereby enhancing the robot's interaction with its environment and task performance.

The remainder of the paper is arranged as follows: Section II presents a review of the related works. Section III provides details of the sensor design and fabrication. Section IV outlines the methods used to do perception using the proposed sensing finger. Section V contains details regarding experiments and results. Finally, Section VI concludes the study and presents the future course of work.

II. Related Works

Tactile sensation is crucial for humans to manipulate objects effectively [17]. To replicate this capability in robotics, researchers have developed tactile sensors that endow robots with human-like tactility [18]. These sensors interact with the robot's environment, capturing key physical quantities such as pressure, force, displacement, surface texture, and depth [19]. By processing these measurements, robots can infer essential

perceptual information necessary for dexterous manipulation [20].

Various sensing mechanisms have been explored to achieve tactile perception in robots. Vision-based tactile sensors employ cameras and computer vision techniques to capture deformations or contact patterns [21], [22]. Vibration-based sensors utilize accelerometers and acoustic transducers to analyze surface interactions [23]. Piezoelectric tactile sensors leverage the piezoelectric effect to convert mechanical stress into electrical signals [24]. Capacitive and resistive sensors measure variations in capacitance and resistance, respectively, to detect contact and force distribution [25], [26]. Additionally, electromagnetic and triboelectric sensors exploit electromagnetic induction and triboelectric effects to generate tactile feedback [27], [28].

Photonics and optical tactile sensors are rapidly emerging, detecting mechanical deformation through changes in light properties [29]. Fiber Bragg Grating (FBG) sensors, a key type, are valued for applications like tactile sensing, surgery, structural monitoring, and biosensing [30], [31]. While highly sensitive and stable, they face challenges with low spatial resolution and complex fabrication.

Among the various categories of tactile sensors, vision-based tactile sensors exhibit superior performance, offering higher spatial resolution, consistency, and robustness [6]. These sensors typically use a camera to capture the deformation of an elastomer gel pad (EGP) during interactions with the environment. GelSight [32] is a well-established vision-based tactile sensor, known for its high spatial resolution and ability to reconstruct fine contact geometries [33], [34]. Recent applications of GelSight sensors include shear and slip measurement [35], texture recognition [36], and liquid property estimation [37].

Despite their inherent advantages, vision-based tactile sensors frequently face a bottleneck in operating at low bandwidth, often in the range of dozens (30 Hz), thereby impeding the execution of dynamic and high-frequency manipulation tasks [38]. To surmount this challenge, it becomes evident that achieving high spatial resolution alone is inadequate; a concurrent emphasis on high bandwidth becomes imperative.

Vibration/acoustic-based tactile sensors are another prominent category among tactile sensors [39]. They offer high-bandwidth sensing suitable for capturing high-frequency signals [40]. Some examples of vibration-based tactile sensors developed for robotic manipulation include [41]–[43]. Despite their high bandwidth, these sensors lack the spatial resolution necessary for complex manipulation [38].

To enhance robots' capabilities in complex manipulation tasks, a tactile sensor that provides both high-resolution and high-bandwidth sensing is required. While individual studies have explored either high-resolution or high-bandwidth tactile sensing, there is limited work on seamlessly integrating both modalities within a single sensor. This paper introduces a novel sensing finger, *VibTac*, which addresses this gap by offering both high-resolution and high-bandwidth sensing in a unified hardware solution.

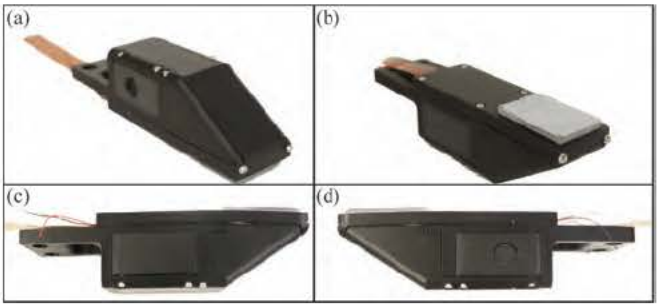

Fig. 3: **Design**: Physical prototype of the *VibTac* finger. (a) Isometric view (top), (b) Isometric view (bottom), (c) Side view (right), (d) Side view (left).

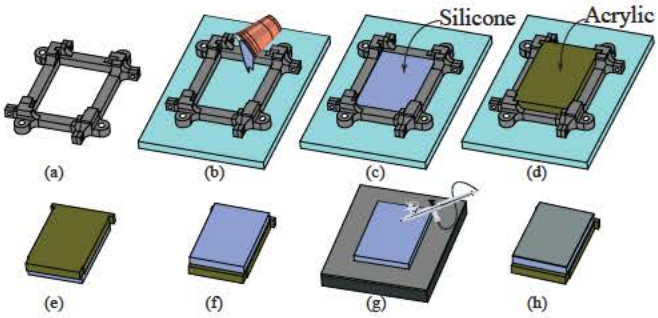

Fig. 4: **EGP Fabrication Process**:(a) Assembling the mold frame, (b) Attaching the back acrylic plate and pouring silicone, (c) Fully pouring silicone, (d) Applying a coat of primer to the acrylic and attaching it to the silicone, ensuring no air bubbles are trapped. After placing the acrylic on top, the silicone is left to cure either at room temperature for 24 hours or in an oven for 30 minutes at 60°C. Once cured, the silicone-acrylic cartridge is demolded. (e) Demolded cartridge with acrylic on top (green) and silicone on the bottom. (f) Demolded cartridge showing the silicone on top, (g) Placing the cartridge in the painting guide and applying reflective coating, (h) Final elastomer gel pad (EGP). A video demonstration of the full fabrication process is included in supplementary material.

III. Design and Fabrication

This section outlines the design of the proposed tactile sensor and its fabrication methods. Fig. 2 presents an exploded view of the sensor design, detailing all the main components. Fig. 3 displays the actual prototype of the sensor. The proposed *VibTac* finger incorporates two main sensing modalities: vision-based tactile sensing and acoustic-based tactile sensing. Each component will be discussed individually.

A. Vision-Based Tactile Sensing

In the following discussion, we delve into the detailed aspects of the vision-based tactile components.

Elastomer Gel Pad (EGP) The proposed sensor features two sensing modes: Vision-based tactile sensing and acoustic-based tactile sensing. Vision-based tactile sensing, utilizing an EGP, provides localized tactile information crucial for the sensor's functionality.

During EGP fabrication, two main considerations are addressed. First, the EGP must be clear without any stains to ensure a clear vision. Second, proper bonding with the acrylic pad is crucial to prevent air trapping between the EGP and acrylic sheet, which could lead to sensing errors. Fig. 4 illustrates the fabrication process, involving custom frame printing, joining, attaching to a clear acrylic sheet, pouring

silicone mixture, curing, and applying a Lambertian coating for tactile sensing enhancement. The coating blocks external light, enabling the camera to capture localized information. The supplementary video included with the manuscript shows the description of the pad fabrication procedure.

Air can become trapped in the silicone during mixing, making it unsuitable for use. To eliminate these bubbles, the silicone should be degassed in a vacuum chamber. Additionally, pouring the silicone slowly into the mold helps prevent air from being trapped between layers or between the silicone and acrylic. Pressing the top acrylic firmly onto the silicone is also crucial to remove any excess air and ensure a proper seal.

Lighting Lighting plays a crucial role in vision-based tactile sensors, especially when employing photometric stereo for tactile sensing. To achieve better results, directional lighting is used to compute deformation depth. While the ideal lighting angle is 120° [44], we arrange three surface-mounted LEDs from the 'Chanzon' brand in a compact strip at 90° for a slender, manipulation-friendly sensor design (Fig. 2). Gray and diffuser filters are currently attached to each LED array—the 'VViViD Air-Tint Dark Air-Release Vinyl Wrap Film' gray filter minimizes reflections in the EGP caused by the acrylic, and the '3M Diffuser 3635-70' ensures consistent illumination and softens bright light spots. It is important to note that the reflections minimized by the gray filter differ from those produced by the gel pad's coating. Reflections from the acrylic are undesirable, as they result from LED light bouncing off the acrylic surface. In contrast, the reflections generated by the coating are beneficial, enabling the computation of essential gradients for perception.

Camera The camera is the key component of the proposed sensor for vision-based tactile sensing. We use the camera to capture pad deformations and extract relevant tactile information. Specifically, we employ the Raspberry Pi Zero standard focus camera, with dimensions $12 \times 12 \times 5$ mm. The small size of the camera also ensures the overall compactness of the robotic finger. The video feed from the camera is streamed using `mjpg-streamer`. The stream provides a feed at a resolution of 640 × 480 pixels, delivering it at 60 Frames Per Second (FPS) and at 90 FPS for a resolution of 320 × 240 pixels.

B. Acoustic-Based Tactile Sensing

The primary goal is to enhance the bandwidth of vision-based tactile sensing by incorporating acoustic sensing. Two distinct sensing elements are employed for a broader sensing bandwidth: an ADXL345 accelerometer captures small-frequency vibrations within 0-200Hz, while a MAX9814 electret microphone amplifier is able to capture high-frequency signals up to 44 kHz. The ADXL345, with a sensing sensitivity of up to 16g (where g is the acceleration due to gravity), is compact and cost-effective, suitable for most microcontroller boards. The MAX9814, equipped with auto gain for noise filtering, offers sound with significantly less noise and is both compact and affordable, presenting an effective solution for our sensor.

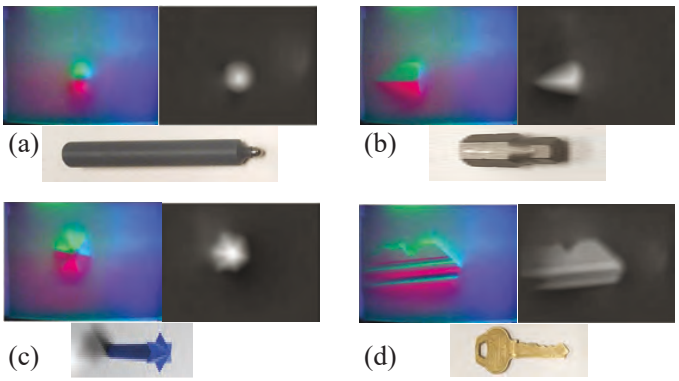

Fig. 5: **Depth estimation results from vision-based tactile sensing:** In each part, the image on the left displays the RGB output from the sensor, while the image on the right shows the corresponding depth estimation. (a) Metal ball, (b) Screwdriver head, (c) 3D printed part, (d) Key.

C. Assembly

All parts of the sensor are manufactured using the Bambu Lab X-1 Carbon 3D printer. After printing, heat set inserts are attached for screwing in different assembly components.

Silicone is cast using the method discussed above, as shown in Fig. 4. During casting, the bottom and top acrylic sheets are cleaned with a nanofiber cloth to prevent imprints on the casted pad. After pouring, the mold is gently tapped to remove large air bubbles. It is worth noting that some small air bubbles will remain even after tapping and will be removed later during the curing process. Finally, the casted pad is demolded and painted.

The accelerometer and microphone are attached to their holders, mirrors are attached to the mirror holder, and LED strips with the acrylic+EGP cartridge are in the camera holder. The camera is then attached, and all components are joined on the skeleton using M-2 screws. The video included with the manuscript shows the full procedure of the sensor assembly. The overall cost of the *VibTac* sensing finger is less than $50.

IV. Perception

The sensor in this paper has two sensing modalities: vision-based tactile sensing and vibration/acoustic-based tactile sensing. Through this confluence, we aim to leverage the benefits of both. Vision-based tactile sensors provide high-resolution data but have limited bandwidth, while vibration/acoustic sensing offers high bandwidth without high resolution. The perception methods for vision-based tactile and vibration/acoustic sensing are different, and this section provides details on the methods used for perception.

A. Vision-Based Tactile Sensing

The main methods for vision-based tactile perception in this project are photometric stereo [45] and Principal Component Analysis (PCA) [46]. Photometric stereo provides 3D geometry information about the object in grip, while PCA is used to estimate the *in-hand* pose.

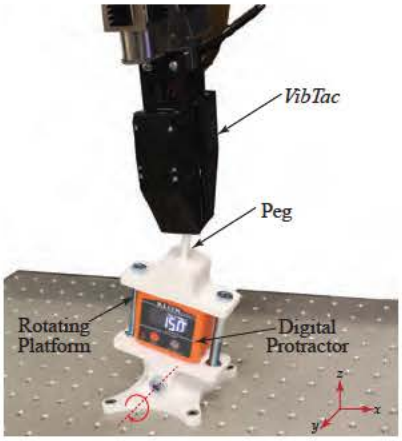

Fig. 6: **Setup to test the accuracy of the PCA method:** In this experiment, the robot was provided with a peg at a known angle (measured using a digital protractor), and the angle estimated by the PCA was then recorded. The results are reported in Table I

1) Photomeric Stereo:

Photometric stereo employs the retrographic method to derive the shape of an object from various lighting variations. The fundamental concept involves representing the object's surface using a height function expressed as $z_i = f(x_i, y_i)$, where x_i and y_i denote the i^{th} spatial coordinates in pixel space. Extracting depth entails two primary steps: Step 1 involves estimating surface gradients $(G^i x, G^i y)$ at spatial location (x_i, y_i) in RGB space using (R_i, G_i, B_i) values. In Step 2, a rapid Poisson solver [47] is employed to spatially integrate surface gradients for achieving complete spatial depth. Once spatial depth is obtained, a point cloud is generated by projecting pixel coordinates x_i, y_i, and z_i as $X_i = (x_i - x_c) * mpp$, $Y_i = (y_i - y_c) * mpp$, and $Z_i = z_i * mpp$, where (x_c, y_c) represents the center coordinates in 2D-pixel space, and mpp denotes the millimeter-per-pixel ratio.

Color-to-Gradient Mapping In the initial phase, our objective is to acquire knowledge of the relationship between R, G, B values at (x, y) and their corresponding surface gradients through the utilization of a Neural Network (NN). This model takes as input the RGB values at a given spatial location and predicts the surface gradients. The optimization process revolves around minimizing the Mean Squared Error (MSE) between the predicted and actual surface gradients.

$$\min_{\theta} \mathcal{L}MSE(\theta) = \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \|g_i - f_\theta(R_i, G_i, B_i)\|_2^2, \quad (1)$$

In this equation, g_i represents the ground-truth surface gradients of the i^{th} data point. These reference values are derived from a calibration ball, which is spherical and has a known diameter. The calibration process involves determining the millimeter-per-pixel (mpp) ratio, creating impressions of a 3D printed sphere, and adjusting the parameters (θ) of the mapping function using the defined optimization objective.

During operation, the trained NN f_θ is responsible for predicting surface gradients. Unlike previous approaches [12], we employ a low-memory footprint neural network for learning the color-to-gradient mapping. This entails utilizing a multilayer perceptron with tanh activation and incorporating three hidden layers, resulting in enhanced performance.

Fast Poisson Solver: Following the gradient computation, a 2D fast Poisson solver [47] is utilized to ascertain depths. This solver utilizes G_x and G_y as inputs, incorporating boundary conditions, to yield relative depths. By leveraging the relative depth information and the mpp ratio, the actual depth is calculated, culminating in the formation of a 3D point cloud. Figure 5 illustrates the tactile depths of diverse objects produced by the *VibTac* sensor, accompanied by their corresponding RGB image.

This approach of extracting depth using image gradients offers several advantages over directly computing depth using a neural network. The latter method introduces significant complexity, as depth prediction is highly sensitive to variations in lighting and surface textures, making it difficult to obtain precise ground truth data for training.

2) Denoising and PCA: To obtain robust PCA results, a stable stream of contact features from the EGP is necessary. We subtract a no-contact background image from the stream, generating a grayscale difference signal. The grayscale frames are dilated and eroded to connect contact regions and denoise. The live grayscale video is then denoised using a first-order autoregressive filter as described by the following equation:

$$Y_t = \alpha X_t + (1 - \alpha) Y_{t-1}, \quad 0 \leq \alpha \leq 1 \quad (2)$$

X_t represents the raw grayscale frame at time t, and Y_t is the corresponding denoised frame. The coefficients α balance the current frame and the previous denoised frame.

Otsu's adaptive binarization [48] is utilized from OpenCV [49] to make the threshold sensitive to contact features and relieve the heterogeneity from light. After isolating the contact region, those thresholded contours are sorted by area; only the largest region is used to compute PCA in this work. The orientation of objects can be obtained by averaging the angles of the first principal component over a few seconds.

To evaluate the accuracy of the PCA method in estimating the *in-hand* angle of the object, we conducted experiments using the setup shown in Fig. 6. A custom 3D-printed platform was designed to rotate the peg around the y-axis, with a digital protractor installed to record the ground truth angle. The *VibTac* finger was mounted on the WSG gripper of the *Franka Panda* robot arm, with the fingers aligned to the z-axis. During the experiment, the peg was rotated at angles from -30° to 30° in 5° increments, with each angle calibrated using the digital protractor. For each angle, the peg was grasped 20 times, and the angle was estimated using PCA orientation. The mean angle and standard deviation results are presented in Table I .

B. Vibration /Acoustic Based Tactile Sensing

1) Accelerometer:

Calibration For accelerometer calibration, we use the method proposed in [50], taking advantage of static conditions where the accelerometer output aligns with gravitational acceleration. The calibration model includes the offset and scaling

TABLE I: Statistical Data of PCA Accuracy

Angle	-30	-25	-20	-15	-10	-5	0	5	10	15	20	25	30
Mean	-30.99	-26.44	-20.99	-16.16	-10.48	-5.06	0.08	6.30	10.36	15.55	20.54	25.48	29.32
Std Dev	1.57	1.07	1.36	0.86	1.00	0.94	0.64	0.52	0.96	1.93	3.19	3.04	2.08

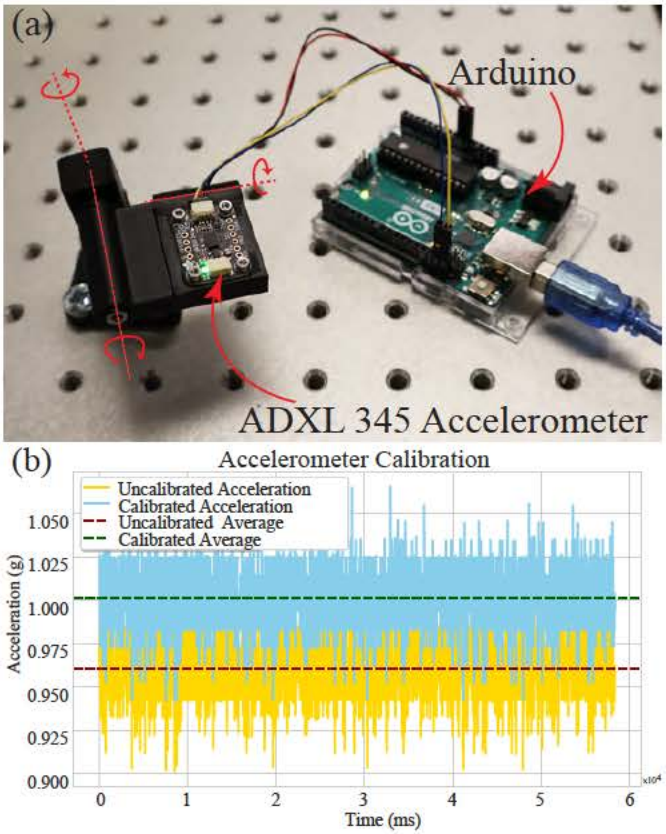

Fig. 7: **Accelerometer Calibration:** (a) The setup used to calibrate the accelerometer, which is attached to a 3D-printed platform that can rotate freely along all three axes (highlighted in red). The accelerometer is connected to an Arduino that records the data during calibration. (b) Calibration results showing both uncalibrated and calibrated outputs. The calibration did not significantly shift the average since the accelerometer came pre-calibrated from the manufacturer.

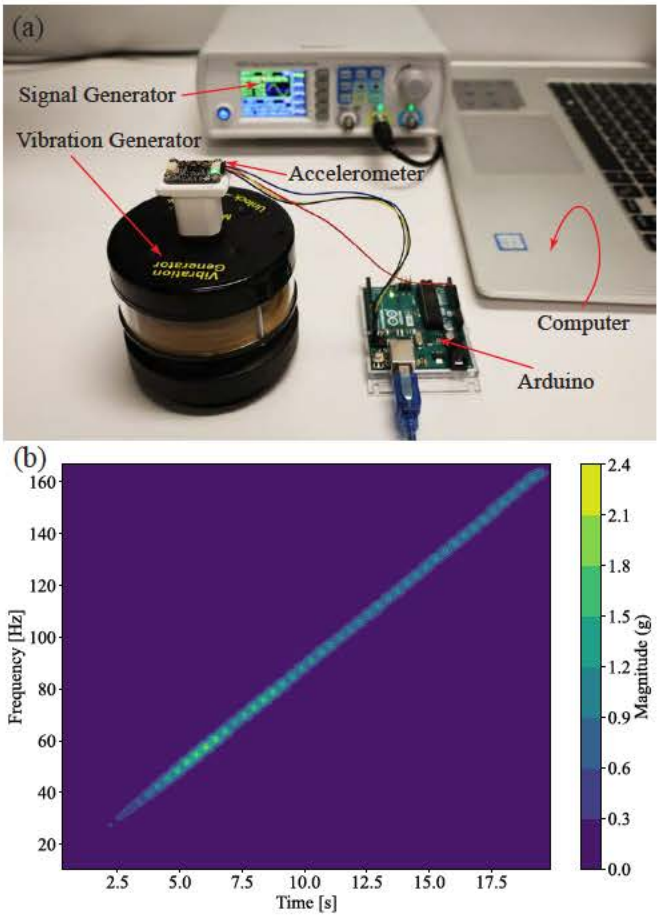

Fig. 8: **Frequency Characterization of the Accelerometer:** (a) Setup used to characterize the accelerometer's frequency response. The accelerometer was mounted on a vibration generator connected to a signal generator, which excited it with a sweep signal ranging from 10 Hz to 160 Hz over 20 seconds. (b) STFT plot of the accelerometer output, showing a clear linear increase in frequency.

values for individual axes, along with factors for cross-axis symmetry, expressed as:

$$\mathbf{C} = \mathbf{F}(\mathbf{A} - \mathbf{B}) \tag{3}$$

where

$$\mathbf{F} = \begin{bmatrix} F_{xx} & F_{xy} & F_{xz} \\ F_{yx} & F_{yy} & F_{yz} \\ F_{zx} & F_{zy} & F_{zz} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} B_x \\ B_y \\ B_z \end{bmatrix} \tag{4}$$

Here, $\mathbf{C}$ is the calibrated value, $\mathbf{F}$ is the scaling factor in each axis, $\mathbf{A}$ is the current accelerometer values, and $\mathbf{B}$ is the bias vector. Fig. 7 shows the calibration setup and results. From the figure, the average acceleration value becomes 1g, as expected as only gravity affects the accelerometer. Calibration involves placing the accelerometer in various orientations using the setup in Fig. 7(a). Fig. 7(b) shows the calibrated and uncalibrated accelerometer output. As the accelerometer works in 3D Cartesian space with gravity, the magnitude of acceleration should be g. As the measurement contains some noise, the reading is averaged for a better view. After calibration, the average output from the accelerometer is g. It is important to note that in our case, calibration does not result in a significant offset due to the high quality of the accelerometer and its accurate factory calibration. However, this may not hold true for all accelerometers, as many can exhibit larger offsets following calibration.

Frequency Characterization We next characterize the accelerometer's frequency response using the setup in Fig. 8, where the accelerometer is mounted on a vibration generator and excited with a sweep signal ranging from 10 Hz to 160 Hz over 20 seconds. A Short-Time Fourier Transform (STFT) is then applied to the recorded data to analyze the frequency components. The resulting STFT plot shows a clear linear increase in frequencies from approximately 10 Hz to 160 Hz.

2) Microphone:

Spectral Analysis

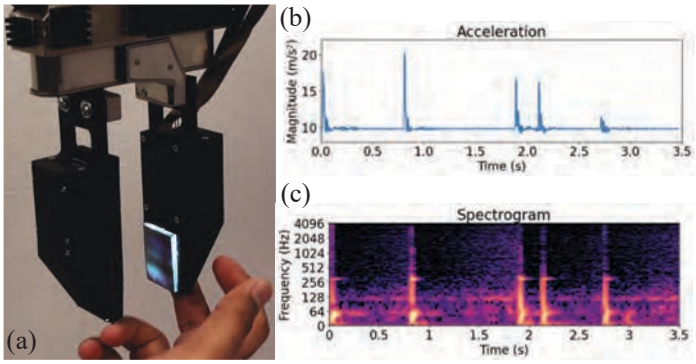

Fig. 9: **Spectral Analysis**: Results from the spectral analysis of microphone data. A Short-Time Fourier Transform (STFT) was applied to detect vibration signatures, generating a decibel spectrogram. (a) The *VibTac* finger was tapped five times to test the pipeline. (b) Accelerometer output during sensor poking. (c) Spectrogram of the microphone output during tapping. Peaks in both (b) and (c) correspond to touch events, enabling the extraction of vibration signatures for various tasks.

To detect vibration signatures from the microphone, we perform Short-time Fourier Transform (STFT) to obtain a decibel (dB) spectrogram using Librosa [51] package with the default setting. The spectrogram provides information in the frequency and time domain, which can be further analyzed using computer vision algorithms. Fig. 9 shows the raw output and corresponding spectrogram when human fingers apply an external mechanical perturbation five times. Since the working sampling rate is 8000 Hz, it can be observed that the maximum frequency in the spectrogram is 4000 Hz due to Nyquist frequency [52].

V. Experiments

A. Data Collection

We attach the designed sensor to a WSG-32 gripper mounted on a *Franka Panda* robot for experiments (Fig. 1). We select objects that produce distinct clicking sounds when they are inserted into their respective sockets (Fig. 10). For data collection, the sockets are fixed onto an aluminum fixture plate in front of the robot, while objects are placed between the WSG gripper's fingers, positioned 6–12 cm above the sockets. The *Franka Panda* robot then moves downward to insert the objects, stopping at a predefined target. During each insertion, both acceleration and microphone data are recorded. A total of 500 insertion trials are conducted for each object, with additional control data collected, capturing only movement noise without object insertion.

B. Dataset and Preprocessing

The collected dataset includes eight objects: Ethernet cable, big buckle, car seat belt, snap button, glue stick, secure pen, buckle, plane seat belt, and background movement noise data. Fig.10 shows all the objects included in the dataset. Each dataset contains an audio file and an acceleration file collected simultaneously during insertion.

We keep complete and valid data and clean garbled characters, randomly keeping 400 pairs of balanced data for each category, that means each category contains nearly the same number of data points, and 800 pairs of blank background data.

Each valid acceleration data contains around 2000 steps of discrete three-dimensional acceleration, and we calculate the magnitude using the $L-2$ norm. Each recorded audio contains around 4 seconds of data in real-time. We remove the center amplitude offset and over-amplification in the audio to standardize the audio, upsample to 22 kHz, apply a short-time Fourier transform (STFT), and obtain dB amplitude spectrograms.

Finally, the corresponding audio and acceleration files are matched, and the two types of data are aligned based on visual pattern inspection. The acceleration data are standardized, and the spectrograms are normalized to zero and one.

C. Deep Learning Model for Multimodal Classification

The objective of the classification is to classify acoustic signals made by different insertions, like a "click" sound. Also, a click detection model is needed to generalize the click. Recurrent neural network (RNN) with the module of gated recurrent unit (GRU) [53] and Long Short-Term Memory (LSTM) [54] is widely used for acoustic and time sequence tasks [55], [56]. One advantage of using RNN is that it can take sequence data in the form of segments and capture the relevance among them. This feature is particularly suitable for online tasks that require contextual input.

The classification algorithm used in this work is a multi-modality RNN model, which accepts both audio spectrogram and acceleration input. The architecture is shown in Fig 11. The aligned data sequences are first divided into seven valid segments in total, each around 0.5 seconds. The corresponding standardization and normalization are applied to each segment individually. Each acceleration segment vector with a length of 143 is input into a double-layer multilayer perceptron (MLP) encoder with ReLU [57] activation and output size of 8. The audio spectrogram is fed into a modified ResNet-18 [58] network as a convolutional neural network (CNN) encoder. Each segment of the spectrogram is a single-channel input with a size of 1025 x 22. The ResNet directly takes the shape without resizing by PyTorch's implementation [59]; its first layer is changed to accept single-channel input, and its last fully connected layer is removed and outputted a vector of size 512.

Then, the output of encoders is concatenated into a vector of size 520. The compacted representation is accepted as the input of an LSTM network. The last LSTM module's output hidden states are linked to a double-layer MLP classifier with 50% dropout [60] to make the prediction.

D. Hyperparameter Grid Search and Modality Ablation Study

To determine the optimal model hyperparameters and modality combination while improving search efficiency, we design a sub-dataset with 160 samples selected from each category. The dataset is randomly split into two halves: one for the training set and the other for the test set. The model is trained for 30 epochs with a batch size of 64. The cross-entropy loss function is used for multi-class classification, and the Adam optimizer [61] is employed.

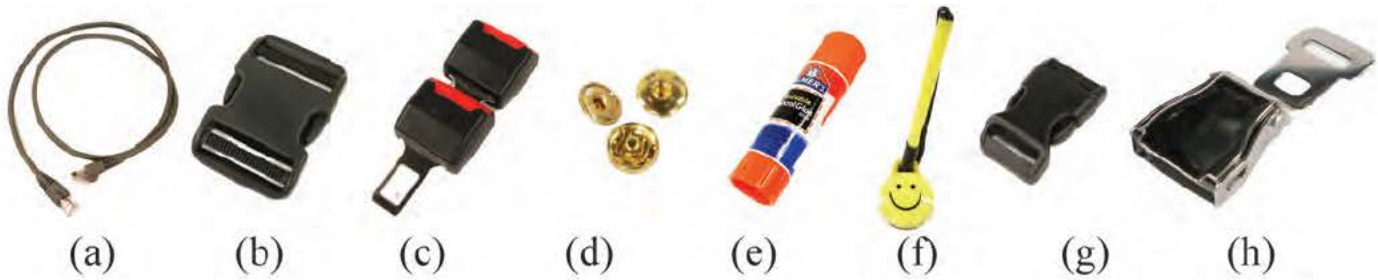

Fig. 10: Objects included in the dataset for the insertion and "click" detection experiment. From left to right: (a) Ethernet cable, (b) Big-buckle, (c) Car seat belt, (d) Snap button, (e) Glue stick, (f) Secure pen, (g) Buckle, (h) Plane seat belt.

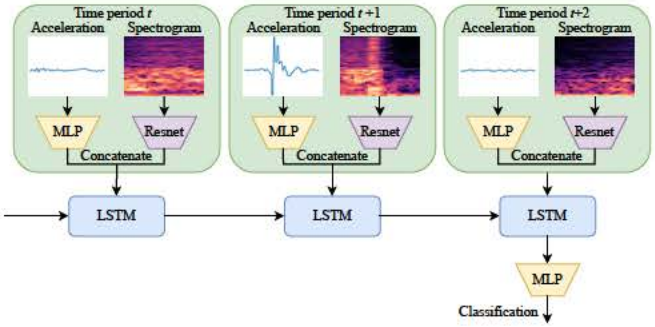

Fig. 11: Architecture of the multimodal RNN used in this paper for detecting vibration signatures.

We perform a grid search over hyperparameters and modalities:

- **Learning Rate**: The learning rates tested are $1e^{-3}$, $1e^{-4}$, and $1e^{-5}$.
- **LSTM Hidden Size**: Hidden sizes of 64, 128, and 256 units are tested to capture temporal information.
- **Modality Ablation**: RNN models with single modalities (spectrogram and acceleration) are tested for an ablation study.

During training, the weights achieving the highest accuracy on the test data are saved for evaluation. The grid search results show that a learning rate of $1e^{-4}$ and an LSTM hidden size of 128 yield the highest accuracy. Additionally, we observe that the multi-modality RNN outperforms the single-modality models, verifying that both modalities contribute to better classification performance.

All experiments are conducted on a system equipped with an Intel Core i9-13900K processor, 128 GB of RAM, and a GeForce RTX 4090 graphics card.

TABLE II: Grid search of multimodal RNN model's hyperparameters and modality ablation study with a sub-training set.

Modality	Learning rate	LSTM hidden size	Accuracy (%)
Spectrogram + Acceleration	$1e^{-5}$	64	79.45
		128	79.66
		256	73.07
	$\mathbf{1e^{-4}}$	64	97.19
		128	**98.23**
		256	96.98
	$1e^{-3}$	64	93.20
		128	94.09
		256	92.03
Spectrogram	$1e^{-4}$	128	97.53
Acceleration	$1e^{-4}$	128	30.60

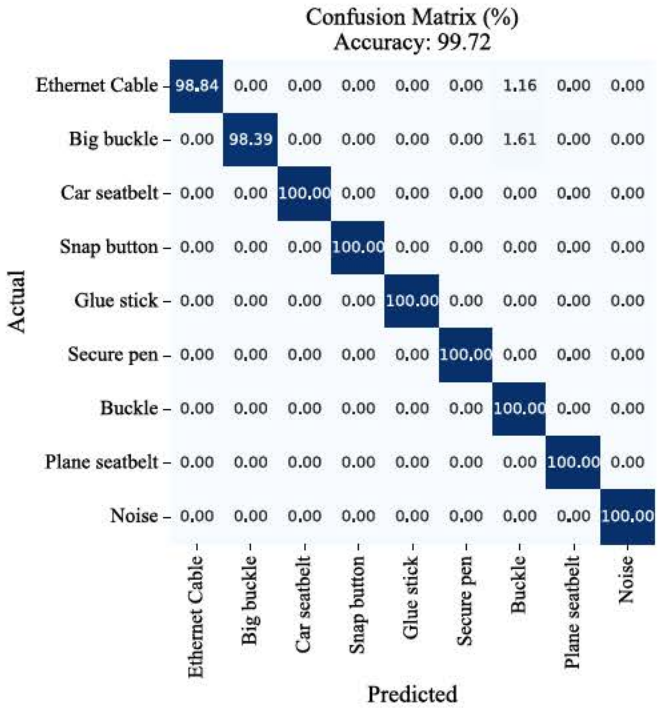

Fig. 12: Confusion matrix (%) illustrating the classification performance of the proposed RNN model, with an overall accuracy of 99.72%. The matrix shows the correct and misclassified predictions across all object categories.

E. Results

1) Classification: For the classification task, the goal is to detect the category of the insertion click. The dataset is combined with 400 samples from each category and the background noise. Twenty percent of the data is split for the test set. The best hyperparameter configuration with a learning rate of $1e^{-4}$ and an LSTM hidden size of 128 is used. The classification model is trained with 30 epochs.

Fig. 12 shows the confusion matrix of the click classification results. It can be observed that the model achieves excellent performance with an overall accuracy of 99.72% in classifying the multi-modal sequence data.

2) Zero-shot Click Detection: For insertion click detection, this work aims to establish this sensor and algorithm's ability to detect common insertion and not only the cases that have been entered and trained in the dataset. We preserve 100 samples of data from each insertion category but leave only one set entirely out of the training set as the test category.

TABLE III: Zero-shot Click Detection Accuracy

Category	Ethernet cable	Big buckle	Car seatbelt	Snap button	Glue stick	Secure pen	Buckle	Plane seatbelt	Average
Accuracy	99.40%	99.52%	99.40%	97.84%	95.67%	93.51%	99.40%	99.40%	98.02%

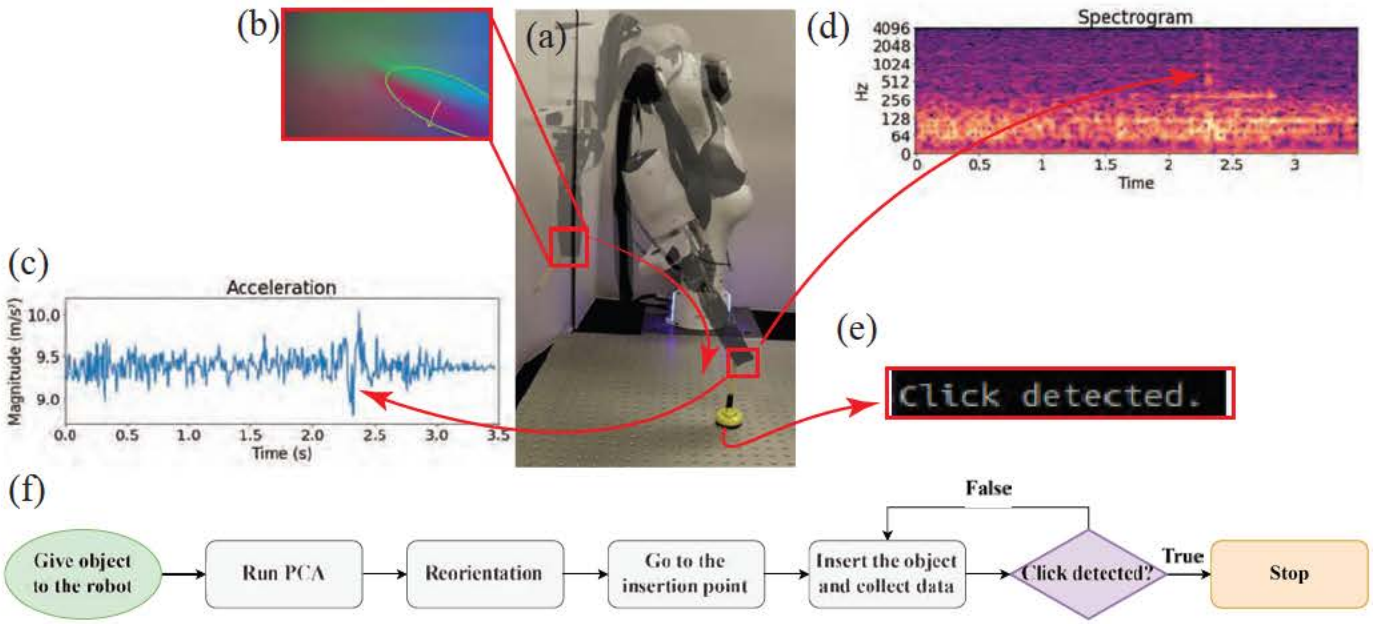

Fig. 13: **Demo**: The complete pipeline for the insertion task using the proposed *VibTac* sensor. (a) The *VibTac* sensing finger mounted on a *Franka Panda* robot during the pen insertion task. (b) Tactile sensing feedback used by the robot to adjust the pen's orientation. (c) Accelerometer output showing a spike upon pen contact. (d) Corresponding spectrogram from the microphone data, detecting the click sound during insertion. (e) The system detects the click and displays "Click detected" on the terminal, signaling the robot to stop. (f) Flow diagram illustrating the task sequence, from object reorientation to insertion and click detection.

The noise data is half separated into the training and test data. Such a dataset ensures that no similar data is available during the training, enabling us to test unseen data exclusively. The learning rate used is $5e^{-4}$, and 40 epochs for each training. The LSTM hidden size is 128.

Table III shows the metrics of the accuracy of detecting the unseen data. The average accuracy of the eight categories is 98.02%, and we find higher accuracy in detecting insertions that emit high volume and high impact. The only accuracy below 95% is the class of secure pen. This is reasonable since its closure mechanism is looser than other classes. It can be inserted easily, making a smaller sound amplitude and impact than others. As a result, it is the least perceptible object for the model.

These results demonstrates our model's capability for zero-shot learning to generalize the click sequence multi-modal data. This generalization reveals our method's plug-and-play potential to handle industrial and daily tasks.

F. Insertion Demo

To demonstrate the overall multi-modal tactile sensing ability of the sensor, we use the sensor to complete various insertion tasks that make distinct "click" sounds. The successful completion of these tasks requires both high-resolution and high-bandwidth tactile sensing, which is achieved by our proposed finger sensor with vision-based tactile modality and acoustic-based modality, respectively.

In the demo, we give objects to the robot at a random orientation and ask it to insert them successfully. For this, the robot first uses vision-based tactile sensing to infer the *in-hand* pose of the object and adjusts the pose so that it can be inserted correctly by aligning the axis of the object to the axis of the insertion point. After this, the robot goes to the insertion target point (which is known). During the insertion stroke, the sensor records accelerometer and microphone data. Meanwhile, the latest data is inputted to the click detection RNN model for online inference. If a click is detected, the control loop of the insertion movement is broken, so the robot stops at the right place. Fig. 13 shows the snapshot of the demo when the robot is inserting the secure pen. It also shows sensor output and various points during the demo.

To reduce the impact of environmental noise and increase online robustness, the input audio volume halves. The click-detection model is pre-trained with 100 samples from each object's data and all noise data with three epochs. This ensures a balanced dataset, as we have 800 data points from noise without clicks and 800 click samples, obtained by selecting 100 samples from each of the eight categories.

It can be seen from the figure that the secure pen is given to the robot at a random angle, and the sensor then uses vision-based tactile (Fig. 13(b)) to estimate the *in-hand* pose of the object and then adjusts the pose and goes to the known insertion point. During insertion, it records the microphone (Fig. 13(d)) and accelerometer (Fig. 13(c)) data in sequence. The recorded segmented data is fed to the model every 0.5 seconds, and the multi-modality RNN model processes the data simultaneously. The script prints "click detected" on the

terminal and stops the robot if it detects a click (Fig. 13(e)). Fig.13(f) presents the flow diagram of the demonstration task. The video demonstration of all the objects included in the dataset is shown in the supplementary video; kindly refer to that.

To demonstrate the necessity of a combination of high resolution and high bandwidth, we conduct an ablation study experiment where we ask the robot to complete the task with only one sensing modality at a time. First, we test without PCA, meaning no vision-based high-resolution tactile sensing. In this case, because the robot has no information regarding the orientation of the object, it is not able to adjust the orientation and fails to insert the object. In the second case, we ask the robot to complete the task without acoustic-based high-bandwidth tactile sensing. In this case, two scenarios emerge. In the first scenario, the robot might be able to insert the object, but due to the fact that there is no acoustic feedback, it will not know when to stop and will hit the insertion with a big impact. This is potentially dangerous and can cause serious damage to the robot hardware. In the second scenario, the robot will not know when to stop and will stop before completing the task. From these experiments, we see that for the successful completion of the insertion task, both sensing modalities are required, and only one sensing modality will not be able to complete the task effectively. A video showing the results from the ablation study is included in the media file.

VI. Conclusion & Discussion

In this study, we present a novel sensing finger, *VibTac*, which integrates high-resolution vision-based tactile sensing with high-bandwidth acoustic-based tactile sensing. We detail the design and fabrication of the sensor components to combine these dual modalities. High-resolution tactile sensing is demonstrated through geometry and orientation extraction using photometric stereo and principal component analysis (PCA). For high-bandwidth tactile sensing, we utilize frequency characterization and spectral analysis. We also train and evaluate models for click classification and zero-shot click detection.

To showcase the sensor's integrated functionality, we perform insertion tasks that require pose alignment and produce distinct "click" sounds upon completion. Initially, high-resolution vision-based tactile data is used to infer the *in-hand* pose of the object. During insertion, high-bandwidth acoustic/vibration data is employed to detect the "click" in real time, signaling successful task completion and closing the feedback loop. These experiments highlight the sensor's ability to handle manipulation tasks that are challenging with only one type of tactile sensing. An ablation study further confirms that both modalities are necessary for successful task completion, as a single modality alone is insufficient.

Compared to existing *state-of-the-art* tactile sensors, our sensor demonstrates a broader range of applications, excelling in tasks that other sensors cannot perform. For instance, high-resolution vision-based tactile sensors, such as [32], are unable to perform click detection with the same level of precision as our sensor. Similarly, purely acoustic-based sensors, like those in [41], [43], lack the capability to accurately estimate the *in-hand angle*—a task our sensor handles effectively due to its integrated high-resolution tactile sensing. The multimodal nature of our sensor, combining both tactile and acoustic sensing, offers a significant advantage over existing technologies. This integration enables it to tackle tasks requiring both modalities, delivering enhanced versatility in applications that demand simultaneous tactile and acoustic precision, which current *state-of-the-art* systems cannot achieve independently.

We conduct ablation studies to evaluate the performance of our sensor and assess the contribution of individual sensing modalities—vision and acoustic. We attempt the insertion task using only one modality at a time. In the first scenario, when vision-based tactile sensing is absent, the robot is unable to adjust its pose and cannot complete the insertion. In the second scenario, without acoustic-based tactile sensing, the robot adjusts its in-hand pose but fails to insert the object correctly due to the lack of feedback during the insertion process. Additionally, we investigate the effects of using the microphone and accelerometer individually, as well as in combination. The results show that combining both sensors significantly improves performance, achieving a classification accuracy of 98%. These findings are summarized in Table II.

Despite the impressive performance of the sensor, we observe certain limitations. First, the classification model's accuracy degrades when the insertion speed and acceleration deviate significantly from the training data. This highlights the need for a more diverse dataset to improve generalization. Second, environmental noise levels fluctuate across different testing scenarios, affecting the robustness of acoustic-based tactile sensing. To address this, future work will expand the dataset and incorporate noise-filtering techniques.

Third, the acoustic response of the sensor varies depending on whether an external force is applied to the gel pad or the sensor's rigid frame, as the gel pad dampens higher-frequency vibrations. This may introduce inconsistencies in signal interpretation.

Finally, precise alignment between the object in the robot's hand and the insertion point remains critical for task success. This suggests the potential benefit of reinforcement learning-based insertion strategies for improved adaptability.

Acknowledgment

This work was supported in part by National Science Foundation (NSF) under Award 2423068, and in part by the United States Department of Agriculture (USDA) under Award 2023-67021-39072 and Award 2024-67021-42878.

References

[1] Matthew T Mason. Toward robotic manipulation. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:1–28, 2018.

[2] Brian T Quinn, Chad Carlson, Werner Doyle, Sydney S Cash, Orrin Devinsky, Charles Spence, Eric Halgren, and Thomas Thesen. Intracranial cortical responses during visual–tactile integration in humans. *Journal of Neuroscience*, 34(1):171–181, 2014.

[3] Pedro Silva Girão, Pedro Miguel Pinto Ramos, Octavian Postolache, and José Miguel Dias Pereira. Tactile sensors for robotic applications. *Measurement*, 46(3):1257–1271, 2013.

[4] Qiang Li, Oliver Kroemer, Zhe Su, Filipe Fernandes Veiga, Mohsen Kaboli, and Helge Joachim Ritter. A review of tactile information: Perception and action through touch. *IEEE Transactions on Robotics*, 36(6):1619–1634, 2020.

[5] Roland S Johansson and J Randall Flanagan. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience*, 10(5):345–359, 2009.

[6] Umer Hameed Shah, Rajkumar Muthusamy, Dongming Gan, Yahya Zweiri, and Lakmal Seneviratne. On the design and development of vision-based tactile sensors. *Journal of Intelligent & Robotic Systems*, 102(4):1–27, 2021.

[7] Yusaku Maeda, Kei Tanimoto, Kenichi Sasayama, and Hidekuni Takao. Neural-network-based tactile perception system using ultrahigh-resolution tactile sensor. *IEEE Transactions on Haptics*, 2023.

[8] Jinhui Zhang, Haimin Yao, Jiaying Mo, Songyue Chen, Yu Xie, Shenglin Ma, Rui Chen, Tao Luo, Weisong Ling, Lifeng Qin, et al. Finger-inspired rigid-soft hybrid tactile sensor with superior sensitivity at high frequency. *Nature communications*, 13(1):5076, 2022.

[9] Dale Purves, George J Augustine, David Fitzpatrick, Lawrence C Katz, Anthony-Samuel LaMantia, James O McNamara, S Mark Williams, et al. Mechanoreceptors specialized to receive tactile information. In *Neuroscience*, volume 9. Sinauer Associates Sunderland, MA, 2001.

[10] André Fiebig. The perception of acoustic environments and how humans form overall noise assessments. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 259, pages 8229–8244. Institute of Noise Control Engineering, 2019.

[11] Yu She, Shaoxiong Wang, Siyuan Dong, Neha Sunil, Alberto Rodriguez, and Edward Adelson. Cable manipulation with a tactile-reactive gripper. *The International Journal of Robotics Research*, 40(12-14):1385–1401, 2021.

[12] Branden Romero, Filipe Veiga, and Edward Adelson. Soft, round, high resolution tactile fingertip sensors for dexterous robotic manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4796–4802. IEEE, 2020.

[13] Chenxi Xiao, Aaron Benjamin Woeppel, Gina Marie Clepper, Shengjie Gao, Shujia Xu, Johannes F. Rueschen, Daniel Kruse, Wenzhuo Wu, Hong Z. Tan, Thomas Low, Stephen P. Beaudoin, Bryan W. Boudouris, William G. Haris, and Juan P. Wachs. Tactile and chemical sensing with haptic feedback for a telepresence explosive ordnance disposal robot. *IEEE Transactions on Robotics*, 39(5):3368–3381, 2023.

[14] Sylvia Hanke, Judith Petri, and Diethelm Johannsmann. Partial slip in mesoscale contacts: Dependence on contact size. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 88(3):032408, 2013.

[15] Raul Fernandez, Ismael Payo, Andres S Vazquez, and Jonathan Becedas. Micro-vibration-based slip detection in tactile force sensors. *Sensors*, 14(1):709–730, 2014.

[16] Jae S Son, Eduardo A Monteverde, and Robert D Howe. A tactile sensor for localizing transient events in manipulation. In *Proceedings of the 1994 IEEE International Conference on Robotics and Automation*, pages 471–476. IEEE, 1994.

[17] R.S. Johansson and J.R. Flanagan. 6.05 - tactile sensory control of object manipulation in humans. In Richard H. Masland, Thomas D. Albright, Thomas D. Albright, Richard H. Masland, Peter Dallos, Donata Oertel, Stuart Firestein, Gary K. Beauchamp, M. Catherine Bushnell, Allan I. Basbaum, Jon H. Kaas, and Esther P. Gardner, editors, *The Senses: A Comprehensive Reference*, pages 67–86. Academic Press, New York, 2008.

[18] Ravinder S Dahiya, Giorgio Metta, Maurizio Valle, and Giulio Sandini. Tactile sensing—from humans to humanoids. *IEEE transactions on robotics*, 26(1):1–20, 2009.

[19] Cheng Chi, Xuguang Sun, Ning Xue, Tong Li, and Chang Liu. Recent progress in technologies for tactile sensors. *Sensors*, 18(4):948, 2018.

[20] Akihiko Yamaguchi and Christopher G Atkeson. Recent progress in tactile sensing and sensors for robotic manipulation: can we turn tactile sensing into vision? *Advanced Robotics*, 33(14):661–673, 2019.

[21] Shixin Zhang, Zixi Chen, Yuan Gao, Weiwei Wan, Jianhua Shan, Hongxiang Xue, Fuchun Sun, Yiyong Yang, and Bin Fang. Hardware technology of vision-based tactile sensor: A review. *IEEE Sensors Journal*, 22(22):21410–21427, 2022.

[22] Sheeraz Athar, Gaurav Patel, Zhengtong Xu, Qiang Qiu, and Yu She. Vistac towards a unified multi-modal sensing finger for robotic manipulation. *IEEE Sensors Journal*, 2023.

[23] Kyungseo Park, Hyunwoo Yuk, M Yang, Junhwi Cho, Hyosang Lee, and Jung Kim. A biomimetic elastomeric robot skin using electrical impedance and acoustic tomography for tactile sensing. *Science Robotics*, 7(67):eabm7187, 2022.

[24] Wenzhuo Wu, Xiaonan Wen, and Zhong Lin Wang. Taxel-addressable matrix of vertical-nanowire piezotronic transistors for active and adaptive tactile imaging. *Science*, 340(6135):952–957, 2013.

[25] Perla Maiolino, Marco Maggiali, Giorgio Cannata, Giorgio Metta, and Lorenzo Natale. A flexible and robust large scale capacitive tactile system for robots. *IEEE Sensors Journal*, 13(10):3910–3917, 2013.

[26] Yuan Zhu, Yang Liu, Yunna Sun, Yanxin Zhang, and Guifu Ding. Recent advances in resistive sensor technology for tactile perception: A review. *IEEE sensors journal*, 22(16):15635–15649, 2022.

[27] Guo Yao, Liang Xu, Xiaowen Cheng, Yangyang Li, Xin Huang, Wei Guo, Shaoyu Liu, Zhong Lin Wang, and Hao Wu. Bioinspired triboelectric nanogenerators as self-powered electronic skin for robotic tactile sensing. *Advanced Functional Materials*, 30(6):1907312, 2020.

[28] Shipeng Xie, Yuanfei Zhang, Hao Zhang, and Minghe Jin. Development of triaxis electromagnetic tactile sensor with adjustable sensitivity and measurement range for robot manipulation. *IEEE Transactions on instrumentation and Measurement*, 71:1–9, 2022.

[29] Ni Yao and Shipeng Wang. Recent progress of optical tactile sensors: A review. *Optics & Laser Technology*, 176:111040, 2024.

[30] Wang Peng, Bing Huang, Xuanxuan Huang, Han Song, and Qingxi Liao. A flexible and stretchable photonic crystal sensor for biosensing and tactile sensing. *Heliyon*, 8(11):e11697, 2022.

[31] Guan Lu, Zhihui Shen, Ting Cai, and Yiming Xu. Research on fbg tactile sensing shape recognition based on convolutional neural network. *Sensors*, 24(13):4087, 2024.

[32] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[33] Alexander C Abad and Anuradha Ranasinghe. Visuotactile sensors with emphasis on gelsight sensor: A review. *IEEE Sensors Journal*, 20(14):7628–7638, 2020.

[34] Micah K Johnson and Edward H Adelson. Retrographic sensing for the measurement of surface texture and shape. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1070–1077. IEEE, 2009.

[35] Wenzhen Yuan, Rui Li, Mandayam A Srinivasan, and Edward H Adelson. Measurement of shear and slip with a gelsight tactile sensor. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 304–311. IEEE, 2015.

[36] Rui Li and Edward H. Adelson. Sensing and recognizing surface textures using a gelsight sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[37] Hung-Jui Huang, Xiaofeng Guo, and Wenzhen Yuan. Understanding dynamic tactile sensing for liquid property estimation. *arXiv preprint arXiv:2205.08771*, 2022.

[38] Liang Zou, Chang Ge, Z Jane Wang, Edmond Cretu, and Xiaoou Li. Novel tactile sensor technology and smart tactile sensing systems: A review. *Sensors*, 17(11):2653, 2017.

[39] Mohamed Benali-Khoudja, Moustapha Hafez, J-M Alexandre, Abderrahmane Kheddar, and Vincent Moreau. Vital: A new low-cost vibrotactile display system. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 1, pages 721–726. IEEE, 2004.

[40] Khaled S Ramadan, Dan Sameoto, and Sthephane Evoy. A review of piezoelectric polymers as functional materials for electromechanical transducers. *Smart Materials and Structures*, 23(3):033001, 2014.

[41] DeaGyu Kim, Zhijian Hao, Jun Ueda, and Azadeh Ansari. A 5 mg micro-bristle-bot fabricated by two-photon lithography. *Journal of Micromechanics and Microengineering*, 29(10):105006, 2019.

[42] Yuichi Kurita, Yamato Sueda, Takaaki Ishikawa, Minoru Hattori, Hiroyuki Sawada, Hiroyuki Egi, Hideki Ohdan, Jun Ueda, and Toshio Tsuji. Surgical grasping forceps with enhanced sensorimotor capability via the stochastic resonance effect. *IEEE/ASME Transactions on Mechatronics*, 21(6):2624–2634, 2016.

[43] Timothy McPherson and Jun Ueda. A force and displacement self-sensing piezoelectric mri-compatible tweezer end effector with an on-site calibration procedure. *IEEE/ASME Transactions on Mechatronics*, 19(2):755–764, 2013.

[44] Shaoxiong Wang, Yu She, Branden Romero, and Edward Adelson. Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6468–6475. IEEE, 2021.

[45] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980.

[46] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[47] J. Doerner. Fast poisson reconstruction in python.

[48] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[49] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[50] Iuri Frosio, Federico Pedersini, and N Alberto Borghese. Autocalibration of mems accelerometers. *IEEE Transactions on Instrumentation and Measurement*, 58(6):2034–2041, 2008.

[51] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, 2015.

[52] Harry Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928.

[53] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[54] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.

[55] Hasim Sak, Andrew W Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014.

[56] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.

[57] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[59] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[60] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

VII. Biography Section

Sheeraz Athar earned his Bachelor of Technology (B.Tech.) in Mechanical Engineering from Aligarh Muslim University, India, in 2019. Additionally, he obtained his Master of Philosophy (M.Phil.) in Mechanical Engineering from The Hong Kong University of Science and Technology in 2021. Presently, he is engaged in his Ph.D. studies at Purdue University's Edwardson School of Industrial Engineering.

Xinwei Zhang graduated with a Bachelor of Engineering (B.E) in Mechanical Engineering from Changzhou University, China, in 2018. He furthered his education by earning a Master of Science (MS) in Mechanical Engineering from Boston University in 2020. Currently, he is pursuing his Ph.D. at the Edwardson Scholl of Industrial Engineering at Purdue University..

Jun Ueda is a Professor at the George W. Woodruff School of Mechanical Engineering, Georgia Tech. He received his B.S., M.S., and Ph.D. from Kyoto University, Japan. Before joining Georgia Tech in 2008, he worked at Mitsubishi Electric and served as an Assistant Professor at the Nara Institute of Science and Technology. Ueda has held positions at MIT and directed Georgia Tech's Robotics Ph.D. program (2015–2017). He is a Senior Editor for IEEE/ASME Transactions on Mechatronics, with several authored books and conference chair roles. His honors include the IEEE Early Career Award, Advanced Robotics Best Paper Award, and Nagamori Award.

Ye Zhao is an Assistant Professor at the George W. Woodruff School of Mechanical Engineering, Georgia Tech, where he leads the Laboratory for Intelligent Decision and Autonomous Robots. He earned his Ph.D. from The University of Texas at Austin, specializing in robust motion planning for dynamic legged locomotion, and was previously a Postdoctoral Fellow at Harvard. A finalist for the 2021 ICRA Best Paper Award, Zhao serves as an Associate Editor for IEEE-RAS Robotics and Automation Letters and IEEE Control Systems Letters, and has chaired major conferences.

Yu She serves as an assistant professor at Purdue University Edwardson School of Industrial Engineering. Previously, he held a postdoctoral research position at MIT's Computer Science and Artificial Intelligence Laboratory from 2018 to 2021. He obtained his Ph.D. in the Department of Mechanical Engineering from the Ohio State University in 2018.

His research delves into the convergence of mechanical design, sensory perception, and dynamic control, with a focus on areas such as human-safe collaborative robots, soft robotics, and robotic manipulation.