ABSTRACT:

Tactile perception is essential for skilled robotic manipulation, yet current systems are limited by low sensor resolution, incomplete modality integration, and insufficient interpretation of complex tactile signals. Here we show the Superior Tactile Sensor (SuperTac), a biomimetic, multimodal tactile sensor inspired by the multispectral vision of pigeons. SuperTac integrates multispectral imaging (mid-infrared to ultraviolet) with triboelectric and inertial sensing into a single 1-mm-thick light-field-modulated skin composed of conductive polymer, fluorescent, reflective, and supporting layers. The sensor combines pressure-adaptive force sensing with high-resolution (0.00545 mm²/pixel) and high-precision measurements across force (0.06 N accuracy), position (0.4 mm accuracy), temperature (0–90 °C range), proximity (<15 cm range), and vibration (0–60 Hz range). It achieves over 94% accuracy in discriminating texture, material, sliding, collision, and colour. To interpret this rich multimodal data, we developed DOVE, an 8.5B-parameter tactile language model that enables sophisticated understanding of tactile stimuli. This integrated sensing and interpretation framework could bring robotic touch perception closer to human-like capabilities, with potential applications in manufacturing, healthcare, and service robotics.

One-Sentence Summary:

A pigeon-eye-inspired multimodal high-resolution tactile sensor, combined with a tactile language model, allows robots to achieve human-like tactile perception and understanding of their environment.

INTRODUCTION

Touch is a fundamental sensory modality for robotic manipulation[1], human-robot interaction (HRI)[2], and extended reality (XR)[3] applications. As embodied intelligence advances, the demand for sophisticated tactile sensing capabilities has grown exponentially. High-resolution multimodal tactile sensors, capable of detecting fine object details while capturing diverse physical information, have emerged as a critical focus in both academic research and industrial development[4,5].

Electronic skin (e-skin) based tactile sensors initially demonstrated significant potential for multimodal sensing due to their versatile functional materials[6-8]. However, increasing spatial resolution and sensing modalities in e-skin necessitates denser electrode arrays, resulting in signal crosstalk and complex readout circuitry. In contrast, visuotactile sensing has been proposed as an elegant alternative, offering sub-millimeter spatial resolution

through optical imaging while naturally integrating with modern artificial intelligence frameworks, including computer vision[9], deep neural networks[4], and large language models (LLMs)[10,11]. Despite these advantages, extending visuotactile sensing to incorporate multispectral and non-imaging modalities presents significant technical challenges. While traditional visual systems can readily integrate non-visible light sensors, this approach is hindered in visuotactile systems due to constraints imposed by the sensing skin. Although recent advancements have demonstrated bimodal visuotactile sensors capable of simultaneous temperature-force[4] and material-force[12] sensing, most implementations remain confined to the visible (VIS) spectrum (Supplementary Table 1 and 2). Consequently, the development of truly multimodal visuotactile sensors faces two primary obstacles: limitations in sensing skin design and restricted imaging bandwidth.

Inspired by the remarkable multispectral vision of pigeons[13,14], we introduce the Superior Tactile Sensor (SuperTac) (Fig. 1a and 1b, and Supplementary Video 1 and 2), an integrated multimodal high-resolution (0.00545 mm^2/pixel) tactile sensor that combines multispectral imaging (Fig. 1c), triboelectric sensing (Fig. 1d), and inertial measurement (Fig. 1e). At the heart of SuperTac is a miniaturized sensing unit (Supplementary Note 1 and Supplementary Table 3) featuring light field modulation and multispectral imaging capabilities. The sensor employs a transparency-adjustable multilayered sensing skin composed of a poly (3,4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT: PSS)[15] conductive layer, an ultraviolet (UV) ink fluorescent layer, and a silver powder-coated reflective layer. This design enables different functional modes across various spectra through light field modulation. Additionally, an integrated inertial measurement unit (IMU) provides complementary acceleration and posture data. SuperTac achieves comprehensive sensing capabilities, including force, texture, deformation, temperature, sliding, material properties, distance, vibrations, collision detection, and color recognition (Fig. 1f and Supplementary Video 3). A unique feature of the sensor is its adjustable internal air pressure, which allows for dynamic adaptation of the force-sensing range. Through deep learning integration, SuperTac shows exceptional performance: a force measurement accuracy of 0.06 N, position accuracy of 0.4 mm, temperature range of 0-90°C, proximity detection, vibration sensing from 0-60 Hz, and over 94% accuracy in texture, material, sliding, collision, and colour classification. To showcase its practical applications, we

integrated SuperTac into a dexterous robotic hand and developed DOVE, a specialised tactile language model. DOVE accurately interprets tactile information from manipulated objects, indicating the sensor's potential for advanced HRI and robotic manipulation tasks (Fig. 1g). This integrated approach achieves unprecedented resolution and functionality compared to existing solutions [4,7,9,12,16-31] (Fig. 1h).

MAIN TEXT

Bio-inspired Design of the Multimodal Tactile Sensor

The vertebrate retina contains specialised photoreceptors - rods and cones - with cones enabling colour vision. Unlike humans, pigeons possess an additional type of cone cell sensitive to ultraviolet wavelengths[32], along with specialised retinal molecules for non-imaging perception, such as magnetic field detection[33]. This enhanced visual system enables pigeons to process complex environmental information more comprehensively. Drawing inspiration from these capabilities, SuperTac combines multispectral imaging with triboelectric and inertial sensing to expand the perceptual capabilities of visuotactile sensors. Based on this design, through a single touch, the sensor can obtain information about an object's shape, texture, colour, temperature, and material, as well as the force during contact.

Structural Design and Sensing Mechanism

Visuotactile sensing, which utilises vision for tactile perception[34], has become increasingly valuable for robotic grasping[35] and manipulation[36], particularly given its compatibility with the foundation model frameworks, such as the vision-language-action (VLA) model[37]. Traditional visuotactile sensors typically consist of sensing skin, imaging, and lighting modules. In contrast, SuperTac introduces an innovative design that integrates multispectral imaging, triboelectric signal acquisition, IMU signal acquisition, and lighting modules into a unified multimodal sensing system, significantly enhancing both functionality and integration. This integrated design enables comprehensive environmental interaction through multiple sensing modalities (Fig. 2a). The system can simultaneously detect force, texture, deformation, temperature, material properties, proximity, sliding,

pose, vibration, and colour (Supplementary Table 1 and 2), providing a detailed multisensory representation of physical interactions.

The sensor's design combines multiple functional elements (Fig. 2b). The core innovative part is an adaptive transparency sensing skin coupled with a multimodal sensing system capable of precise spectral band detection, triboelectric signal acquisition, and IMU-based motion sensing. To capture triboelectric signals, we developed a transparent conductive layer based on PEDOT: PSS integrated into the sensing skin. The design also incorporates an IMU for orientation and acceleration sensing. These components are compactly integrated into a four-layer printed circuit board (PCB) implementation with a radius of 16 mm, housing the multispectral imaging, triboelectric, IMU signal acquisition, and lighting modules (Supplementary Note 2 and Supplementary Fig. 1, 2, and 3).

Sensing skin: The selection and structure of sensing skin materials are optimized to enhance SuperTac's functionalities (Supplementary Note 3). The skin comprises four layers: a conductive layer, a reflective layer, a fluorescent layer, and a supporting layer (Fig. 2b and Supplementary Fig. 4), with a thickness of only 1 mm (Supplementary Fig. 5). The conductive layer, fabricated by screen-printing transparent PEDOT: PSS ink on thermoplastic polyurethane (TPU) thin film, generates triboelectric signals during object contact. PEDOT: PSS provides excellent transparency and conductivity, while TPU offers exceptional stretchability, transparency, and toughness (Supplementary Fig. 6). The combination ensures both film transparency and stable triboelectric signal generation. The electrode adopts a vortex line (PEDOT: PSS) design to provide a uniform signal. Based on the triboelectric mechanism (Supplementary Note 4), the conductive layer generates distinct electrical signals upon contact with objects of varying electronegativities, enabling material type discrimination and proximity sensing (Supplementary Fig. 7).

The reflective layer operates similarly to a one-way mirror (Fig. 2c and Supplementary Fig. 8 and 9), of which the transparency is regulated by light intensity on either side: on the bright side, reflected light dominates, rendering the film opaque; on the dark side, transmitted light prevails, making the film transparent. This mechanism allows independent imaging across different wavelengths by controlling the light intensity in specific spectral bands.

The fluorescent layer employs UV light to control marker visibility. These markers, visible

in the UV spectrum but invisible in the near-infrared (NIR) band, enable the sensor to alternate between detection modes with and without markers (Supplementary Fig. 4). This capability allows simultaneous deformation and slide detection without compromising texture detection. When combined with the multispectral imaging system, it captures UV markers and NIR texture information.

The supporting layer is the base substrate of the sensing skin, providing mechanical integrity and structural stability for the entire multilayer assembly. Its main functions are to maintain the overall shape and flexibility of the skin, ensure reliable integration and alignment of the other functional layers (conductive, reflective, and fluorescent), and protect the sensor from mechanical damage during repeated deformations. Additionally, the supporting layer serves as a physical barrier, isolating the functional layers from external contaminants and environmental factors, thereby enhancing the durability and longevity of the sensor. Unlike traditional acrylic-based designs, we employ a silicone-based inflatable support structure. This design offers several advantages: a larger deformation range for detailed object contour representation, an adjustable force-sensing range (0 to 7 N) through internal air pressure control (Supplementary Fig. 10), and improved thermal response due to its thinner profile. Additionally, the silicone inflatable film addresses the limitations of mid-infrared (MIR, 5.5 μm to 14 μm wavelength) temperature sensing, where traditional materials like acrylic and standard glass cannot transmit wavelengths above 5 μm. This eliminates the need for costly, special optical glass while maintaining performance. However, the pneumatic support structure offers advantages such as adjustable pressure sensing and enhanced deformation sensing but poses challenges related to sealing, material aging, and repeatability. To address these issues, we integrated a compact air supply system, replaced latex with durable silicone, and utilized TPU film for improved wear resistance, achieving superior durability and consistent performance over 80,000 tests.

Multimodal sensing system: The multimodal sensing system integrates four modules: multispectral imaging, triboelectric signal acquisition, IMU signal acquisition, and lighting modules (Fig. 2b). The miniaturized multispectral imaging module includes an MIR camera, a CMOS camera with low-pass filtering, and a CMOS camera with bandpass filtering. The system covers four spectral bands: UV (390 nm illumination, 450 nm

fluorescence), VIS (400–700 nm), NIR (940 nm), and MIR (5.5–14 μm) (Supplementary Fig. 11). To prevent cross-talk, tactile mode uses UV fluorescence detection, while visual mode captures external visible light with the UV LED turned off. (Fig. 2d).

MIR Detection: For temperature measurement, we employ an MIR imaging camera (MLX90640) with 24×32 resolution, capable of detecting wavelengths between 5.5 μm and 14 μm and measuring temperatures from -40 °C to 300 °C. This camera captures thermal radiation emitted by objects, enabling precise temperature mapping.

NIR Detection: A CMOS unit paired with a 935-945 nm bandpass filter and lens provides precise NIR signal detection, with filter selection determined by the LED light source wavelength.

VIS and UV Detection: A CMOS unit with a 700 nm low-pass filter and lens covers an imaging range from 350 nm to 1000 nm, encompassing UV, VIS, and NIR spectra. LED lighting adjustment enables selective wavelength detection.

The lighting module is meticulously designed to support both reflective and fluorescent layer functionalities. For fluorescent marker detection, 390 nm LEDs excite the fluorescent layer, revealing marker information. The UV fluorescent markers enable modality switching for deformation, sliding, and texture sensing, offering advantages in 3D reconstruction and sliding detection without relying on strict light control. When deactivated, the fluorescent layer becomes transparent, allowing external color observation (Fig. 2c). For texture sensing, 940 nm LEDs generate a strong internal NIR light source, rendering the thin film opaque and enhancing surface texture detection (Supplementary Fig. 12). This light source also works in conjunction with the NIR detection unit, providing stable illumination for precise signal detection (Supplementary Note 5).

For triboelectric signal acquisition, we use an ADA4505 chip operating at a 1 kHz sampling frequency (Supplementary Table 4). The IMU signal acquisition utilizes MPU6050, capturing three-dimensional orientation angles and acceleration data. This configuration enables comprehensive multimodal sensing while maintaining system compactness and integration, addressing the limitations of traditional visuotactile sensors.

SuperTac demonstrates comprehensive sensing capabilities across multiple spectral bands and sensing modalities (Fig. 2d). In the UV band, fluorescent markers enable precise tracking of sliding and deformation through marker size and displacement measurements

(Supplementary Note 6, Supplementary Table 5, and Supplementary Fig. 13, 14, and 15). The VIS spectrum provides object color information upon contact, while the NIR band captures texture and contact force data. Mid-infrared imaging enables temperature measurement, complemented by triboelectric signals for material identification (Fig. 2e) and proximity sensing (Fig. 2f). Additionally, IMU-based collision and vibration detection further enhance the system's multimodal sensing capabilities.

Performance Characterization

To evaluate force and position sensing capabilities, we developed a testing platform incorporating an ATI Gamma sensor as the ground truth for force measurements (Fig. 3a). The evaluation utilized 48 probe (Supplementary Fig. 16) designs across three geometries (U-shape, V-shape, and polygon), collecting approximately 1,800 datasets per probe (Fig. 3b). A force-sensing neural network (Fig. 3c) was developed based on a UNet architecture[38], with ResNet4839 as the encoder to extract features from RGB deformation images captured by the sensor. A fully connected layer was added to output the resultant force vector, while the UNet decoder generated a deformation mask. The mask was multiplied by the resultant vector to produce a force distribution map. The network was trained and evaluated using 86,440 sets of deformation data collected from 48 probe types (Fig. 3d), with a uniform sampling method employed to ensure comprehensive coverage of the sensor surface and accurately assess its force sensing performance. The dataset was split into 70% for training and 30% for testing. Training was conducted on an NVIDIA A6000 GPU using the L1 loss function and the AdamW optimizer, with a CosineAnnealingLR learning rate scheduler. The network achieves a position detection MSE accuracy of 0.056 mm and a 3D force detection MSE accuracy of 0.0004 N, with an overall position detection precision of around 0.4 mm (Fig. 3e) and a force error distribution of approximately 0.06 N (Fig. 3f), demonstrating robust performance across all probe types and strong generalizability (Supplementary Fig. 17). In addition, we conducted comparative experiments using UV and NIR modalities over 80,000 contact events to evaluate force sensing accuracy. Results showed that NIR consistently outperformed UV markers across all evaluation metrics, confirming its superior accuracy and stability in force sensing tasks (Supplementary Fig. 18). For 3D reconstruction testing,

we not only optimized the distribution of markers in simulations but also evaluated the reconstruction accuracy of different algorithms. Through testing, our proposed method achieved an average root mean square error (RMSE) of 0.0892 and mean absolute error (MAE) of 0.0375 (Supplementary Note 6). For surface characterization, a long short-term memory (LSTM) algorithm (Supplementary Note 7 and Supplementary Fig. 19) processed 150 sets of sliding and non-sliding data, achieving 97% accuracy in sliding detection. Color classification was evaluated across six different colors, achieving 100% accuracy. Texture recognition was tested on six 3D-printed textures (Supplementary Fig. 20) and six common textures (Supplementary Fig. 21), demonstrating 98% accuracy (Fig. 3g and 3j). Additionally, the sensor exhibited robust capabilities in Braille sensing as well as the perception of 0.07 mm-thick hair strands (Supplementary Fig. 22). To verify the accuracy of Braille recognition, we collected 200 samples for each of the 26 Braille letters, achieving a classification accuracy of 100%, which demonstrates the sensor's exceptional texture sensing capabilities (Supplementary Fig. 23).

Temperature detection was validated across a range of 0 to 90°C, limited by the thermal resistance of the TPU film (Supplementary Fig. 24 and 25, Supplementary Videos 4 and 5). After testing, the SuperTac can achieve a temperature sensing accuracy of 0.25 °C after calibration and remains unaffected by ambient temperature variations within the 28-50 °C range. UV-induced heating causes only a minimal surface temperature change of 0.2 °C, ensuring negligible interference with MIR-based temperature measurements (Supplementary Note 8 and Supplementary Fig. 26).

The triboelectric sensing capability of SuperTac was comprehensively evaluated under diverse conditions, including 10 different materials, 7 contact surface geometries, 15 contact speeds, 3 contact angles, and 5 pressure levels (Supplementary Note 9 and Supplementary Fig. 27, 28, and 29). Controlled experiments demonstrated robust classification performance in all situations, achieving 97% accuracy for contact angles, 99% accuracy for pressure levels, 96% accuracy for velocities, and 95% accuracy for contact shapes, with an overall 95% accuracy across all conditions (Fig. 3k). A triboelectric signal acquisition platform was developed (Supplementary Fig. 30) to facilitate detailed signal analysis, and a 3.8-hour durability test revealed consistently stable signal output (Supplementary Fig. 31). Furthermore, by employing advanced signal filtering techniques

and neural network classification, the triboelectric signals enabled proximity sensing within a range of 0-15 cm, depending on the material properties, underscoring the versatility and reliability of SuperTac in diverse sensing applications.

Vibration detection capabilities were validated using a custom platform (Supplementary Fig. 32), demonstrating accurate frequency recognition within the range of 0-60 Hz (Fig. 3i and Supplementary Fig. 33). For collision detection, we analyzed 150 sets of IMU signals from collision and non-collision scenarios, achieving 94% classification accuracy (Fig. 3l and Supplementary Fig. 34).

Integration and Applications

Robotic hand implementation:To demonstrate SuperTac's capabilities, we integrated it into two robotic platforms: a three-finger dexterous hand and a parallel gripper (Supplementary Video 6 and Supplementary Fig. 35 and 36). The dexterous hand features 10 degrees of freedom with servo motor actuation at each joint. SuperTac is mounted in the palm, enabling comprehensive object property sensing during grasping operations. For the parallel gripper configuration, SuperTac is installed on one side, facilitating stable object manipulation through integrated visual detection, contact force sensing, slip detection, and collision detection algorithms.

Multimodal tactile language model:To enable advanced tactile information processing, we developed DOVE (Supplementary Note 10 and Supplementary Fig. 37), a multimodal tactile language model built upon a pretrained LLM (Fig. 5d). DOVE fuses multimodal tactile inputs and language to characterize object properties, reason over tactile differences between object pairs, and infer an object's type and function. Specifically, DOVE can process triboelectric, temperature, color, and texture inputs to generate rich descriptions such as *"yellow, room temperature, with a textured, raised, metallic surface."* (Fig. 5d and Supplementary Video 7) When it receives tactile feedback from two objects, DOVE produces relational reasoning statements, e.g., *"The two objects share similar colors, temperatures, and textures, but differ in material, so they are different."* DOVE also associates tactile impressions with semantic knowledge for reasoning, e.g., *"PET is commonly used for food containers. Its yellow color suggests visibility or citrus-related items. This is likely a beverage bottle used for daily consumption."* To explore the impact

of network structure on the perception capabilities of DOVE, we further investigated the effects of the hidden dimensions and activation functions in the projection layer. Experimental results demonstrated that changes in hidden dimensions had minimal impact on performance, while using the GELU activation function significantly outperformed ReLU, ensuring effective alignment and fusion of multimodal features (Supplementary Note 11 and Supplementary Table 6).

Enhanced human-robot interaction: We further demonstrated the system's HRI capabilities across four experimental scenarios (Fig. 5e, Supplementary Note 12, 13, and 14, Supplementary Table 7, Supplementary Videos 8, 9, 10, 11, and 12, and Supplementary Fig. 38 and 39). In the first scenario, the system identifies and selects a metallic cup with a smooth surface. In the second scenario, the system follows user instructions to locate a cup with specific characteristics—lettering and a rough surface. GPT-4o orchestrates the interaction by directing visual identification and physical interaction with each cup, while DOVE processes the tactile feedback. In the third scenario, DOVE receives a reference object via touch and retrieves another that matches a specified color by reasoning jointly over texture and color cues. In the fourth scenario, DOVE infers cluttered tabletop objects' functions as reusable, recyclable, or general waste based on tactile feedback and generates natural-language justifications for each decision. The system continues evaluation until it finds a matching object or determines that no suitable matches exist.

The integration of comprehensive tactile sensing, language-based interpretation, and visual processing represents a significant advancement toward human-like robotic perception and interaction. By enabling robots to process and respond to multimodal sensory information in a manner akin to human perceptual capabilities, this approach paves the way for more intuitive and effective human-robot collaboration.

CONCLUSIONS

Traditional e-skin-based tactile sensors continue to face significant challenges in resolution, homogeneity, and stability. While visuotactile sensors offer promising solutions through advanced imaging techniques, their multimodal sensing capabilities have been limited by constraints in sensing skin design and imaging bandwidth. Our work addresses these fundamental limitations through a light-field modulated sensing skin combined with

multispectral imaging, enabling high-resolution multimodal sensing. The sensor achieves remarkable performance metrics, including 98% texture detection accuracy, 0.06 N 3D force detection accuracy in the NIR band, 97% sliding detection accuracy in the UV band, and 100% color detection accuracy in the VIS band. By incorporating non-imaging perception inspired by pigeon magnetic field sensing, we further extend the sensor's capabilities to material detection (95% accuracy), collision detection (94% accuracy), and vibration detection (0-60 Hz range), all without compromising imaging quality or introducing electrode crosstalk issues.

The interpretation of heterogeneous tactile information through foundation models presents unique challenges. DOVE, our multimodal tactile language model, addresses these challenges through a unified input representation approach, which enhances scalability and adaptability across diverse sensor configurations. However, this approach reveals important trade-offs. While transforming sequential data into images has proven effective for certain tasks, it may not fully capture the temporal characteristics inherent in tactile signals. Alternative approaches, such as time-series encoders, might better preserve temporal features but reintroduce challenges related to embedding heterogeneity. Striking the optimal balance between scalability and effectiveness remains a crucial area for future research and practical implementation.

Several promising directions emerge for extending SuperTac's capabilities. Miniaturization of the sensor could enable fingertip installation, significantly advancing robotic in-hand manipulation capabilities. Additionally, DOVE's modality-agnostic framework, which converts various input modalities into image representations, could be adapted for different sensor configurations and applications. Future work will focus on advancing low-power decoding chips and exploring highly integrated packaging solutions to further reduce the sensor's size while addressing challenges in heat dissipation and system stability, while also optimizing DOVE across diverse sensor designs and application-specific datasets to enhance its versatility and robustness. These developments aim to bridge the gap between robotic and human-like perception capabilities, paving the way for more intuitive and effective HRI.

METHODS

- **Fabrication of the sensing skin**

The sensing skin was fabricated using a multi-step process (Supplementary Fig. 40, Supplementary Note 15, and Supplementary Table 8). First, transparent silicone was mixed and poured into acrylic molds, which provided a smoother surface finish compared to 3D-printed molds. After heating, the silicone is cured to form the supporting layer. For the fluorescent layer, a scraping method was employed, using a steel mesh as a mask to spread fluorescent ink over the surface. To prevent unevenness caused by ink buildup, an additional layer of transparent silicone was applied using spin-coating. The reflective layer was created by mixing silver powder with transparent silicone, which was then spin-coated onto the fluorescent layer. For the conductive layer, conductive ink was screen-printed onto a TPU surface and heated for 60 minutes to complete the layer. Finally, the conductive layer was attached to the translucent layer, finalizing the sensing skin. While the integration of fluorescent markers introduces additional complexity, the standardized design ensures low cost (less than $1) and high durability. The outer sensing skin, made of TPU film commonly used in automotive and smartphone protective applications, exhibits exceptional wear and corrosion resistance. Fluorescent markers showed no photobleaching after one week of continuous UV exposure, ensuring stability (Supplementary Fig. 41). These features demonstrate a thoughtful balance between functional enhancements and cost-effectiveness.

- **Assembly and connection of SuperTac**

The sensor was designed with a modular structure (Supplementary Note 16 and Supplementary Fig. 42), divided into three sections: upper, middle, and lower. The upper and lower sections were made of aluminum alloy for high heat resistance and mechanical strength, while the middle section was constructed from transparent acrylic to ensure even diffusion of LED light onto the sensing skin. Threaded joints were used to connect the modules, allowing for easy disassembly. To address potential overheating during prolonged full-load operation, the SuperTac system incorporates a detachable magnetic cooling fan powered via contact-based pogo pins and aligned using N52-grade magnets, enabling quick removal for maintenance and effectively reducing the stabilized

temperature by 18.4 °C during extended high-load operation, as demonstrated through time–temperature comparison experiments (Supplementary Note 17 and Supplementary Fig. 43, 44, 45, and 46).

The SuperTac system adopts a USB 3.1 Gen1 protocol for data communication, facilitating robust and high-speed transmission across all sensing and communication modules (Supplementary Note 18 and Supplementary Fig. 47). To ensure stable operation, the system is equipped with an optimized power architecture that supports all modules under full-load conditions, with a maximum power consumption of 4.5 W (Supplementary Note 19 and Supplementary Fig. 48 and 49). These design choices enhance the practicality and scalability of the SuperTac system in real-world applications. In addition, we have designed a UI interface that simultaneously displays signals including mid-infrared, near-infrared, visible light & ultraviolet, triboelectric signals, posture information, and acceleration data (Supplementary Fig. 50).

- **Image classification network design and training**

For image-based tactile inputs, a ResNet18 backbone was cascaded with a multi-layer perceptron (MLP) to extract task-relevant features and perform classification. The model processes batches of 128×128 visuotactile images, generating intermediate feature maps through ResNet, which were further processed via max-pooling and passed through the MLP classifier. The network was trained end-to-end for four tasks: color, texture, temperature, and material classification. Triboelectric signals were filtered to remove high-frequency components and visualized as curves, which were stored as images. The dataset was split into 80% for training, 10% for validation, and 10% for testing. The model was trained using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 128, alongside a step scheduler that reduced the learning rate by 0.9 every 10 validation steps.

- **Sequential signal classification network design and training**

For sequential inputs (e.g., IMU data and visuotactile videos), an LSTM network was employed as the backbone to process the temporal flow of information (Supplementary Fig. 19). Low-dimensional data, such as IMU readings, were processed using a two-layer MLP, while spatial-structural data, such as videos, were processed using a pretrained ResNet18.

The LSTM updated its hidden state sequentially and output task-oriented information, which was passed through an MLP classifier for final prediction. For IMU data, the model was trained end-to-end for collision detection, while for sliding detection, only the LSTM and MLP classifiers were trained. The dataset was split into 80% for training, 10% for validation, and 10% for testing. Training used the AdamW optimizer with a learning rate of 1×10^{-3} and a batch size of 128, alongside a step-based learning rate scheduler. After testing, the classification algorithms based on ResNet and LSTM have a single prediction time within 6 ms, meeting the real-time requirements (Supplementary Note 20 and Supplementary Table 9).

- **Effects of air pressure and object hardness on sensor perception**

We investigated the impact of internal air pressure on the tactile sensing performance of the sensor, focusing on its ability to perceive flexible objects and its accuracy in force sensing, texture recognition, and sliding detection. During testing, five pressure levels (1.2 kPa, 3 kPa, 4 kPa, 6 kPa, and 7 kPa) were selected for force sensing experiments, while three pressure levels (3 kPa, 5 kPa, and 7 kPa) were chosen for texture recognition and sliding detection experiments. Experimental results demonstrated that variations in air pressure had minimal impact on the accuracy of force sensing, texture recognition, and sliding detection. Notably, texture recognition and sliding detection achieved 100% accuracy across all pressure conditions. A slight decrease in force sensing accuracy was observed at high pressure (7 kPa), but it remained within an acceptable range. Overall, the system exhibited stable and reliable performance under varying pressure conditions (Supplementary Note 21 and Supplementary Fig. 51 and 52).

Extensive testing of the SuperTac system was conducted on soft and liquid-containing objects, including probes made of diverse materials (PLA, cloth, plastic, paper, PET, silicone) and flexible or liquid-containing textures. While the softness of objects slightly impacted force sensing accuracy, the performance significantly improved by supplementing the dataset with 500 flexible object samples (Supplementary Note 22 and Supplementary Fig. 53). The system achieved 100% accuracy in texture recognition and sliding detection (Supplementary Fig. 54 and 55). Furthermore, the inflatable structure of

SuperTac demonstrated superior texture and contour sensing capabilities compared to GelSight Mini, highlighting its advantages in handling complex surfaces (Supplementary Fig. 56). Additionally, simulation results using finite element analysis (FEA) revealed that the system maintains reliable contour recognition for objects with elastic moduli above 1 MPa, providing theoretical guidance for practical applications (Supplementary Note 23 and Supplementary Fig. 57).

- **Tactile language model design and training**

To enable comprehensive understanding and reasoning over multimodal tactile data and language, a large tactile language model was trained on a processed dataset integrating color, texture, temperature, and triboelectric data, augmented with synthetic tactile language Q&A pairs (Supplementary Fig. 37 and 39). The training and testing data for the SuperTac system were constructed using tactile data spanning 6 colors, 3 temperature conditions, 10 material types, and 6 surface textures, with multimodal Q&A pairs generated by GPT-4 and rule-based scripts to integrate tactile information with natural language descriptions (Supplementary Note 24). The training involved three stages: encoder pretraining, embedding alignment, and model fine-tuning. Pretrained CLIP models[40] were used to extract image features, with an MLP classifier attached for end-to-end classification. After fine-tuning, the classifiers were removed, and a projection layer was added for embedding alignment. Finally, the projection layer and language backbone (Vicuna[41]) were fine-tuned using LoRA[42]. The total parameters of the four CLIP encoders and language backbone reached 8.6 billion. Training used the AdamW optimizer with a cosine annealing schedule, achieving robust performance across all modalities (Supplementary Note 25 and Supplementary Table 10, 11, and 12).

Supplementary Materials

Supplementary Notes 1 to 25
Supplementary Tables 1 to 13
Supplementary Figs. 1 to 57
Supplementary Videos 1 to 12

References and Notes

1 Bauza, M. *et al.* SimPLE, a visuotactile method learned in simulation to precisely pick, localize, regrasp, and place objects. *Science Robotics* **9**, eadi8808 (2024).

2 Luo, Y. *et al.* Adaptive tactile interaction transfer via digitally embroidered smart gloves. *Nature Communications* **15**, 868 (2024).

3 Sun, Z., Zhu, M., Shan, X. & Lee, C. Augmented tactile-perception and haptic-feedback rings as human-machine interfaces aiming for immersive interactions. *Nature Communications* **13**, 5224 (2022).

4 Abad, A. C., Reid, D. & Ranasinghe, A. HaptiTemp: a next-generation thermosensitive GelSight-like visuotactile sensor. *IEEE Sensors Journal* **22**, 2722-2734 (2022).

5 Zhang, N. *et al.* Soft robotic hand with tactile palm-finger coordination. *Nature Communications* **16**, 2395 (2025).

6 Zhong, D. *et al.* High-speed and large-scale intrinsically stretchable integrated circuits. *Nature* **627**, 313-320 (2024).

7 Qiu, Y. *et al.* Quantitative softness and texture bimodal haptic sensors for robotic clinical feature identification and intelligent picking. *Science Advances* **10**, eadp0348 (2024).

8 Çeliker, H., Dehaene, W. & Myny, K. Multi-project wafers for flexible thin-film electronics by independent foundries. *Nature* **629**, 335-340 (2024).

9 Yuan, W., Dong, S. & Adelson, E. H. Gelsight: high-resolution robot tactile sensors for estimating geometry and force. *Sensors* **17**, 2762 (2017).

10 Yang, F. *et al.* Binding touch to everything: learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 26340-26353 (IEEE, 2024).

11 Fu, L. *et al.* A touch, vision, and language dataset for multimodal alignment. Preprint at https://arxiv.org/abs/2402.13232 (2024).

12 Mu, S. *et al.* A platypus-inspired electro-mechanosensory finger for remote control and tactile sensing. *Nano Energy* **116**, 108790 (2023).

13 Hochstoeger, T. *et al.* The biophysical, molecular, and anatomical landscape of pigeon CRY4: a candidate light-based quantal magnetosensor. *Science Advances* **6**, eabb9110 (2020).

14 Xu, J. *et al.* Magnetic sensitivity of cryptochrome 4 from a migratory songbird. *Nature* **594**, 535-540 (2021).

15 Shi, H., Liu, C., Jiang, Q. & Xu, J. Effective approaches to improve the electrical conductivity of PEDOT: PSS: a review. *Advanced Electronic Materials* **1**, 1500017 (2015).

16 Sun, H., Kuchenbecker, K. J. & Martius, G. A soft thumb-sized vision-based sensor with accurate all-round force perception. *Nature Machine Intelligence* **4**, 135-145 (2022).

17 Hogan, F. R. *et al.* Seeing through your skin: recognizing objects with a novel visuotactile sensor. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* 1218-1227 (IEEE, 2021).

18 Li, S. *et al.* M $^{3} $ Tac: a multispectral multimodal visuotactile sensor with beyond-human sensory capabilities. *IEEE Transactions on Robotics* **40**, 4506-4525 (2024).

19 Park, K. *et al.* A biomimetic elastomeric robot skin using electrical impedance and acoustic tomography for tactile sensing. *Science Robotics* **7**, eabm7187 (2022).

20 Zhang, Y., Kan, Z., Tse, Y. A., Yang, Y. & Wang, M. Y. Fingervision tactile sensor design and slip detection using convolutional lstm network. Preprint at https://arxiv.org/abs/1810.02653 (2018).

21 Ward-Cherrier, B. *et al.* The tactip family: soft optical tactile sensors with 3d-printed biomimetic morphologies. *Soft Robotics* **5**, 216-227 (2018).

22 Ge, J. *et al.* A bimodal soft electronic skin for tactile and touchless interaction in real time. *Nature Communications* **10**, 4405 (2019).

23 Massari, L. *et al.* Functional mimicry of Ruffini receptors with fibre Bragg gratings and deep neural networks enables a bio-inspired large-area tactile-sensitive skin. *Nature Machine Intelligence* **4**, 425-435 (2022).

24 Liu, W. *et al.* Touchless interactive teaching of soft robots through flexible bimodal sensory interfaces. *Nature Communications* **13**, 5030 (2022).

25 Jiang, Y. *et al.* A multifunctional tactile sensory system for robotic intelligent identification and manipulation perception. *Advanced Science* **11**, 2402705 (2024).

26 Mao, Q., Liao, Z., Yuan, J. & Zhu, R. Multimodal tactile sensing fused with vision for dexterous robotic housekeeping. *Nature Communications* **15**, 6871 (2024).

27 Wang, Y. *et al.* Hierarchically patterned self-powered sensors for multifunctional tactile sensing. *Science Advances* **6**, eabb9083 (2020).

28 Cai, M. *et al.* A multifunctional electronic skin based on patterned metal films for tactile sensing with a broad linear response range. *Science Advances* **7**, eabl8313 (2021).

29 Hua, Q. *et al.* Skin-inspired highly stretchable and conformable matrix networks for multifunctional sensing. *Nature Communications* **9**, 244 (2018).

30 Kim, K. *et al.* Extremely durable electrical impedance tomography-based soft and ultrathin wearable e-skin for three-dimensional tactile interfaces. *Science Advances* **10**, eadr1099 (2024).

31 Tian, X. *et al.* High-resolution carbon-based tactile sensor array for dynamic pulse imaging. *Advanced Functional Materials* **34**, 2406022 (2024).
32 Roorda, A. & Williams, D. R. The arrangement of the three cone classes in the living human eye. *Nature* **397**, 520-522 (1999).
33 Mouritsen, H. & Hore, P. J. The magnetic retina: light-dependent and trigeminal magnetoreception in migratory birds. *Current Opinion in Neurobiology* **22**, 343-352 (2012).
34 Li, S. *et al.* When vision meets touch: a contemporary review for visuotactile sensors from the signal processing perspective. *IEEE Journal of Selected Topics in Signal Processing* **18**, 267-287 (2024).
35 Li, S. *et al.* Visual-tactile fusion for transparent object grasping in complex backgrounds. *IEEE Transactions on Robotics* **39**, 3838-3856 (2023).
36 Suresh, S. *et al.* NeuralFeels with neural fields: visuotactile perception for in-hand manipulation. *Science Robotics* **9**, eadl0628 (2024).
37 Brohan, A. *et al.* Rt-2: vision-language-action models transfer web knowledge to robotic control. In *Proceedings of Conference on Robot Learning* 2165-2183 (PMLR, 2023).
38 Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-assisted Intervention* 234-241 (Springer, 2015).
39 He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770-778 (IEEE, 2016).
40 Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning* 8748-8763 (PMLR, 2021).
41 Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the Advances in Neural Information Processing Systems* 46595-46623 (Curran Associates, 2023).
42 Hu, E. J. *et al.* Lora: low-rank adaptation of large language models. Preprint at https://arxiv.org/abs/2106.09685 (2021).

Acknowledgments

Funding:

Acknowledgements: National Key R&D Program of China (2024YFB3816000; W.D.); Shenzhen Science and Technology Program (KJZD20240903100905008; W.D.);

Guangdong Innovative and Entrepreneurial Research Team Program (2021ZT09L197; W.D.); Tsinghua Shenzhen International Graduate School–Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation (SZPR2023005; W.D.); National Natural Science Foundation of China (12472160; Z.X.); Research Grants Council of the Hong Kong Special Administrative Region (RFS2324-1S03, R1017-24F, T42-513/24-R, C7005-23Y, 11211425, 11215722, 11211523; X.Y.); National University of Singapore Presidential Young Professorship Start Up Grant (A-8000380-02-00; C. W.); Ministry of Education, Singapore, Academic Research Fund (A-8002146-00-00; C. W.); Tsinghua–NUS Joint Research Initiative Fund Award (A-8002542-00-00; C. W.).

Author contributions:

W.D. and S.L. conceived the idea and guided the project. S.L., T.W., and J.X. designed the experiments, analyzed the data, and drafted the manuscript. Z.X., C.W., and X.Y. instructed in manuscript writing and experimentation. S.L. and T.W. performed the characterization of the material. S.L. and T.W. conducted functional experiments on the sensor. T.W. and Y.H. designed and implemented classification algorithms. S.L. and J.X. contributed to the mechanical design. Z.Z., H.Z., and Y.Y. conducted theoretical analysis and simulations. S.L., T.W., Q.X., Z.W., S.M., L.Y., X.W., Z.X., C.L., C.W., X.Y., and W.D. revised the manuscript. All authors discussed the results and provided comments on the manuscript.

Competing interests:

The authors declare that they have no competing interests.

Data availability:

The data that support this study are available at https://cloud.tsinghua.edu.cn/d/f6abfcf5845a42018e2a/files/?p=%2FData%2Fdataset.zip.

Software availability:

All experiments were conducted using Python 3.8.20 in a Conda environment. All analyses were performed on Ubuntu 20.04 with 4 NVIDIA RTX A6000 GPUs (CUDA 11.3).

Code availability:

We have open-sourced the codebase for DOVE at https://github.com/wut19/DOVE. Future updates and new releases will also be available at this link.

Figures:

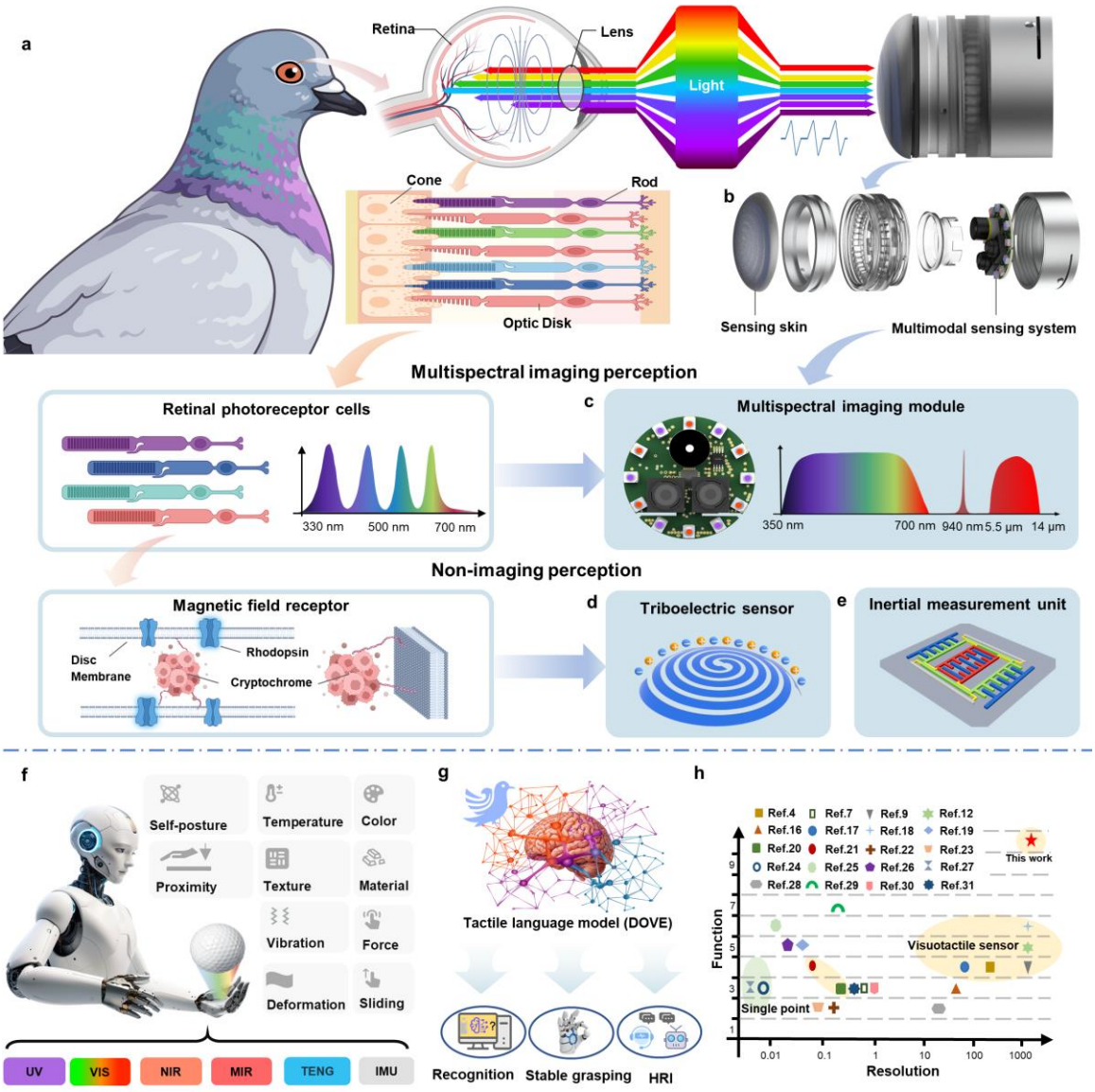


Fig. 1. Overview of the multimodal tactile sensing system. (**a**) The structure of the retina in pigeons includes cones and rods. We draw inspiration from the remarkable multispectral vision along with specialized retinal molecules for non-imaging perception, such as magnetic field detection. (**b**) The overall structure of the sensor comprises a sensing skin and a multimodal sensing system. (**c**) Multispectral imaging systems achieve visible (VIS), ultraviolet (UV), near-infrared (NIR), and mid-infrared (MIR) spectral sensing. (**d**) Triboelectric sensor and (**e**) inertial measurement unit (IMU) to enhance the sensing capability of the tactile sensor. (**f**) SuperTac's demonstration of sensing modalities and functions. Deploying sensors with a manipulator can enable the sensing of ten functions. (**g**) SuperTac combined with the tactile language model (DOVE) can be applied in object recognition, grasping, and HRI. (**h**) Comparison of current mainstream tactile sensors regarding resolution and functionality.

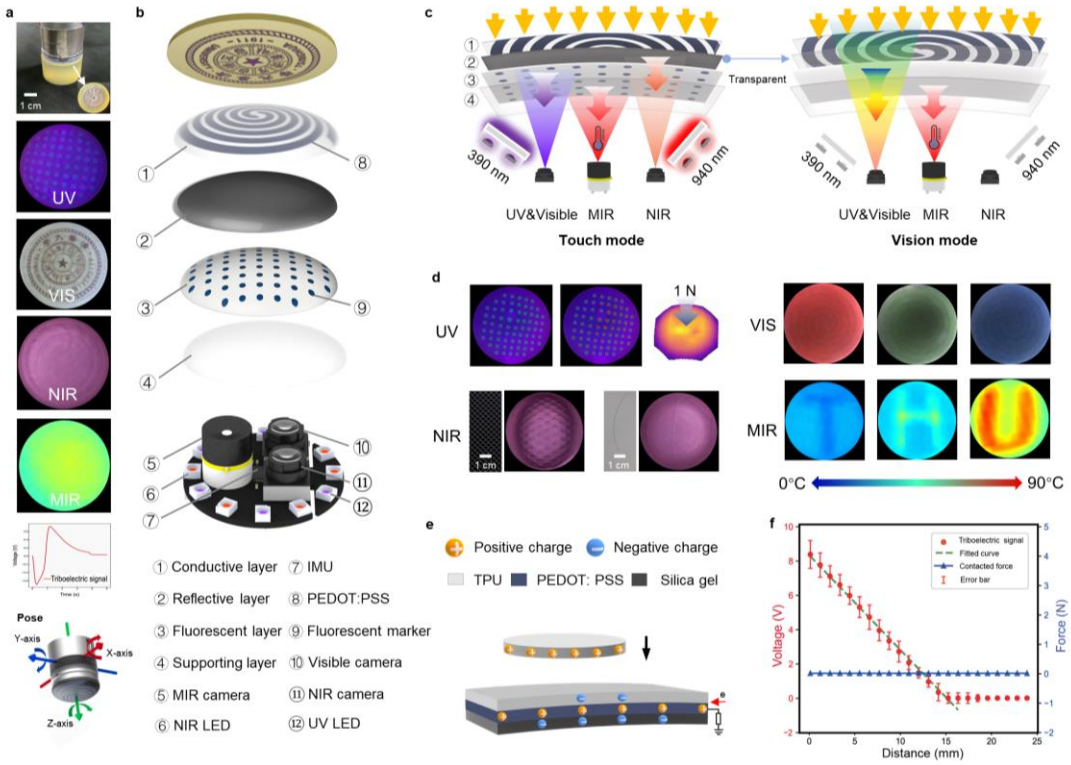


Fig.2. Structural Design and Sensing Mechanism. (**a**) Sensing modalities of SuperTac. (**b**) The structure of sensing skin, imaging module, and lighting module. (**c**) SuperTac is in touch mode when the internal lighting module is turned on, while it is in vision mode when the internal lighting module is turned off. (**d**) The tactile data is captured when the sensor is in contact with the object. In column-first order: the UV image in touch mode; the NIR image in touch mode; the RGB image in vision mode; and the temperature data. (**e**) Triboelectric signal acquisition mechanism. (**f**) Object proximity sensing（Each data point's error bar is based on n=5 independent experimental repetitions, and the error bar represents the maximum and minimum values of the error）.

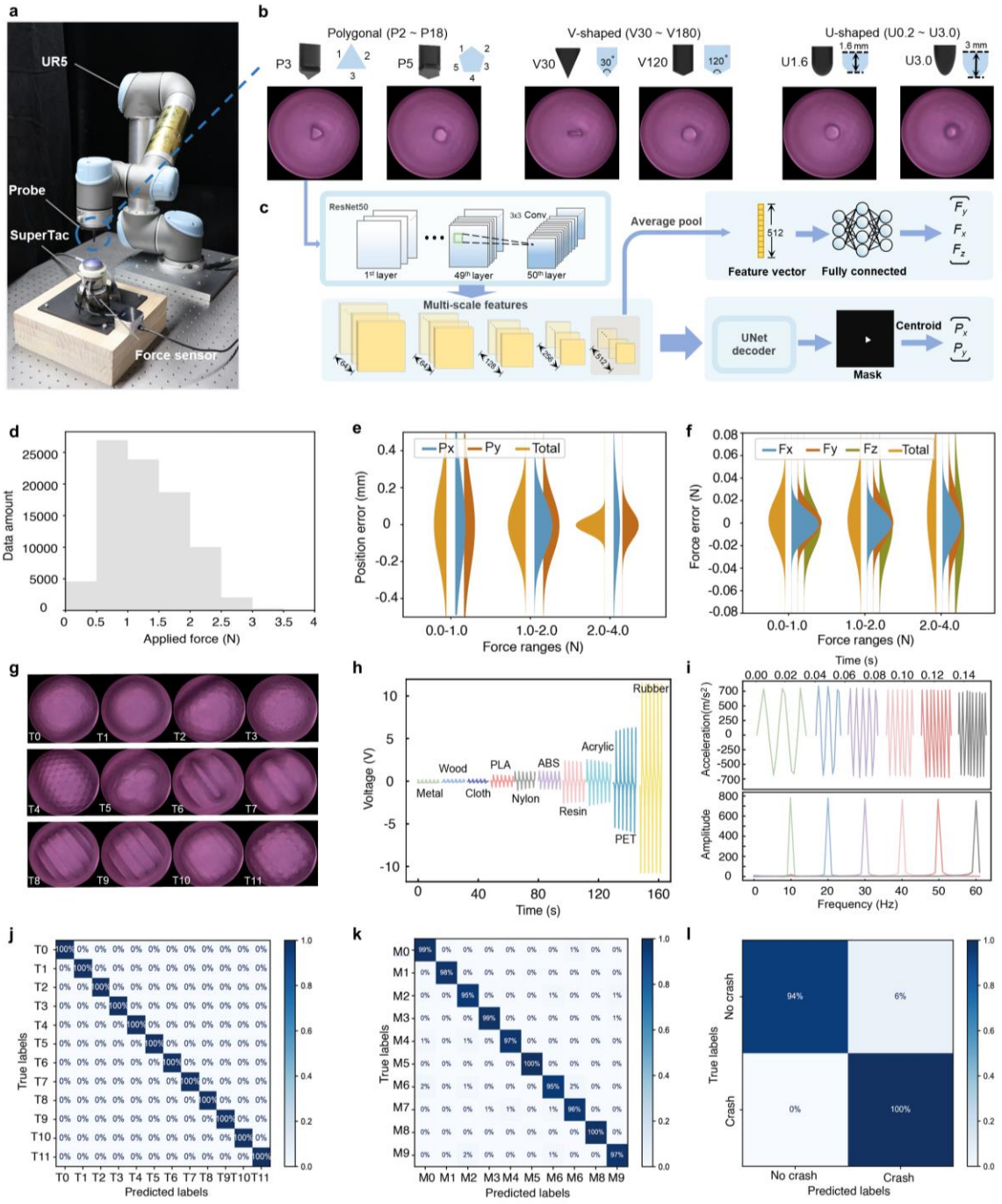


Fig.3 Perception and classification algorithm design. (**a**) Force sensing data acquisition platform. (**b**) We test the force sensing accuracy of 48 probes in U-shape, V-shape, and polygonal shapes. (**c**) Force sensing network. (**d**) In the experiment, we collected 86,440 data sets for contact force distribution. (**e**) Contact position detection accuracy. (**f**) Force sensing accuracy. (**g**) Textures of 12 different surfaces. (**h**) Triboelectric signal of 10 different materials. (**i**) Vibration signals at different frequencies are detected by the SuperTac. (**j**) Texture classification confusion matrix. (**k**) Material classification confusion matrix. (**l**) Collision detection confusion matrix.

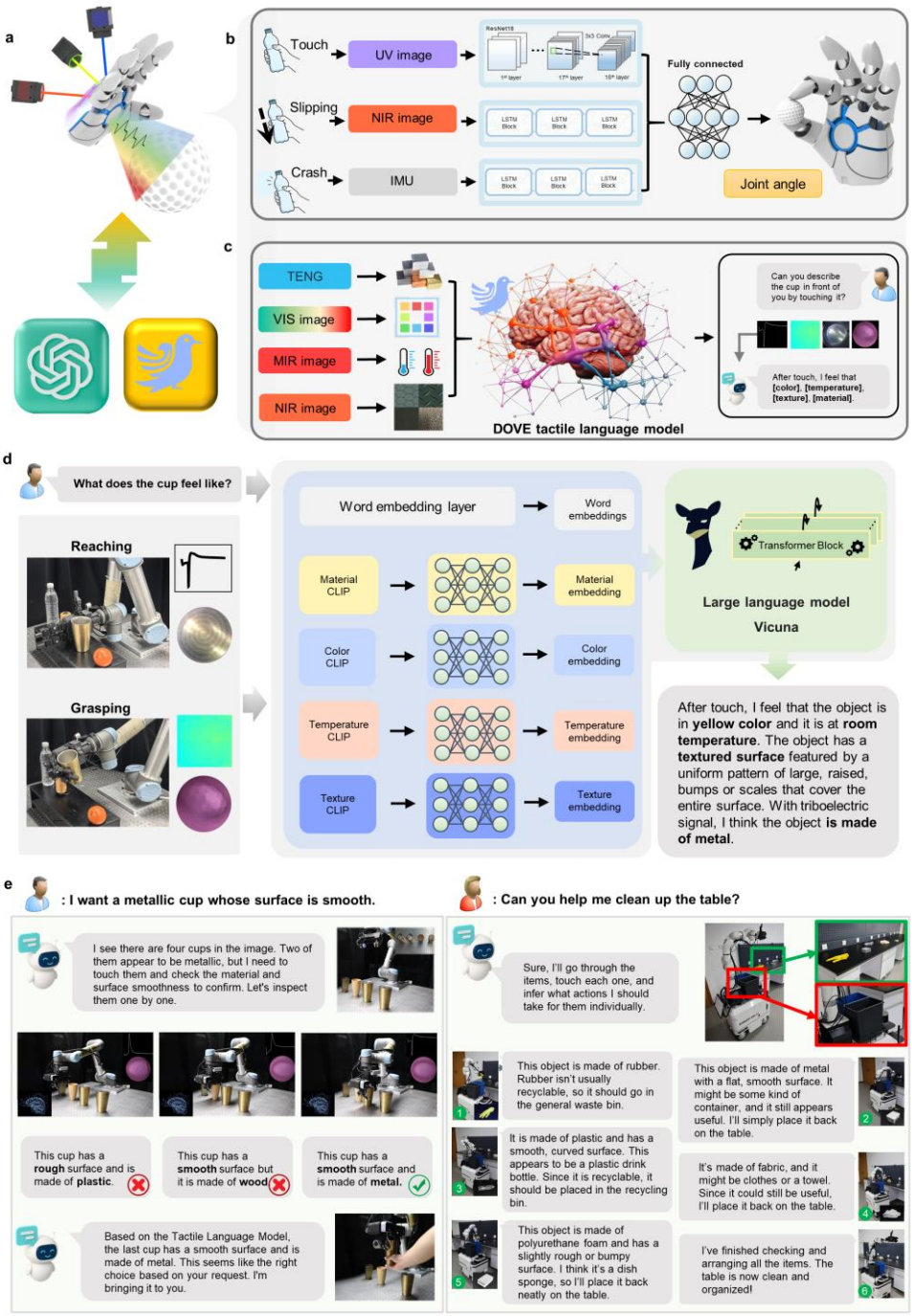
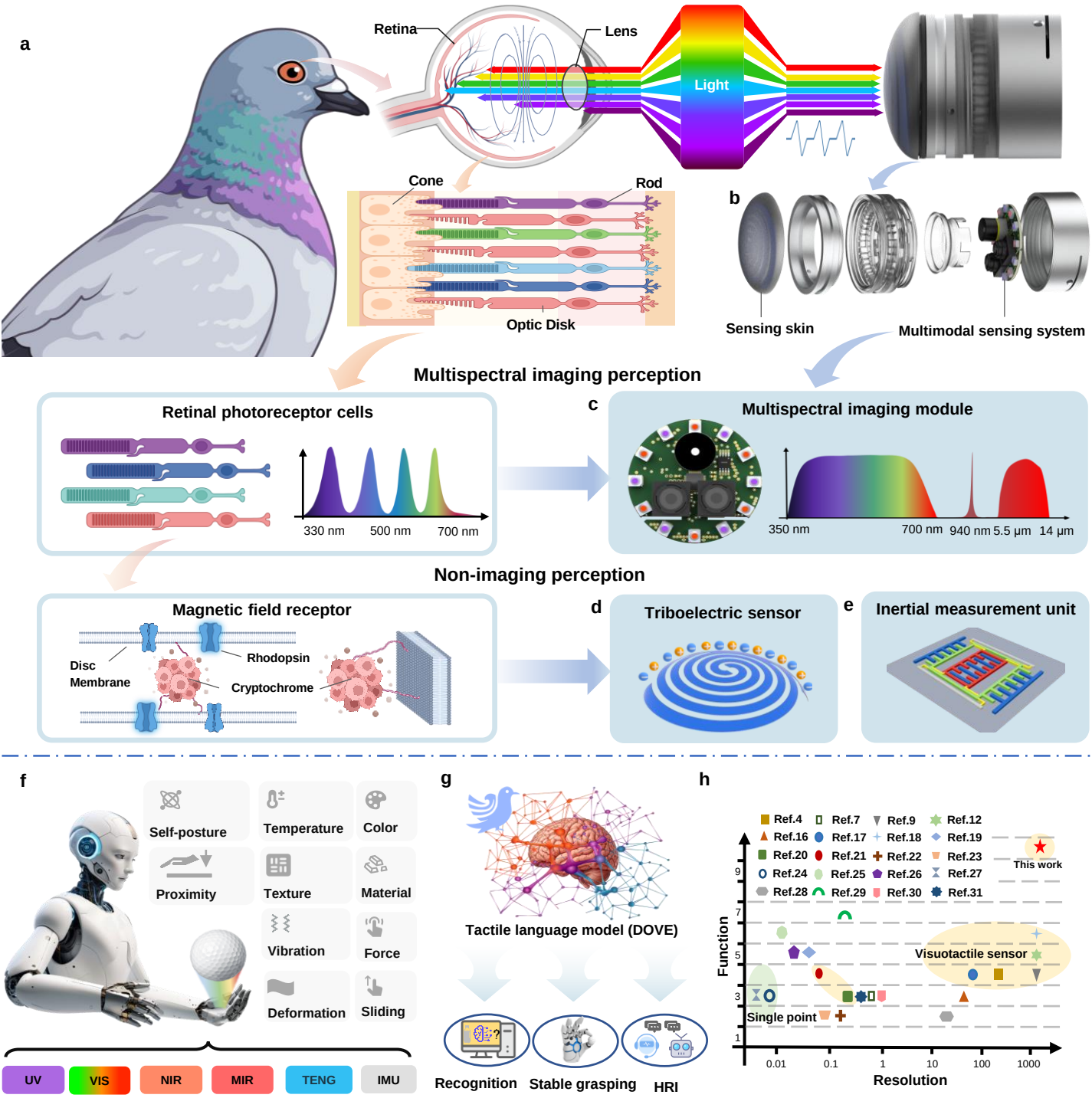
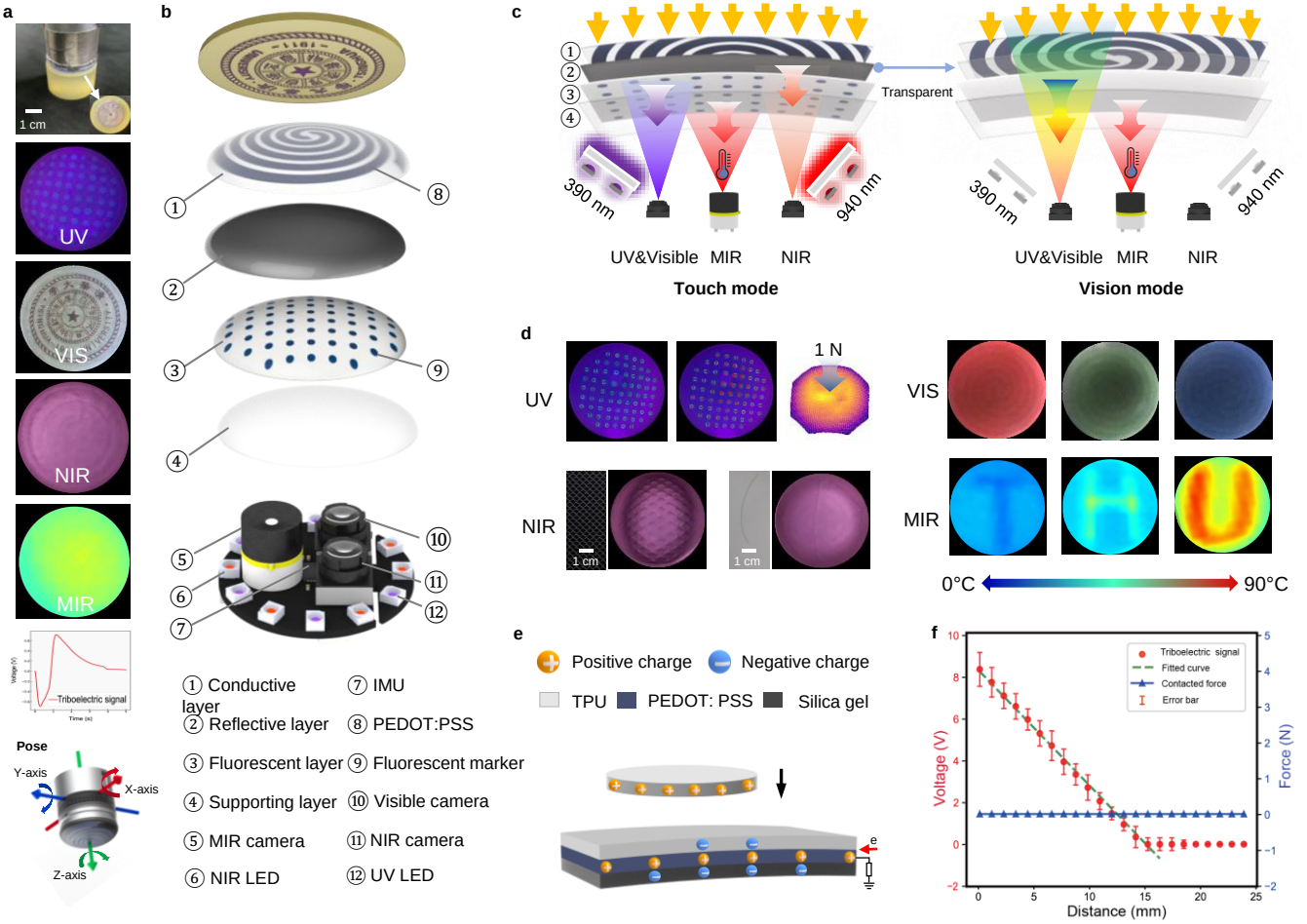
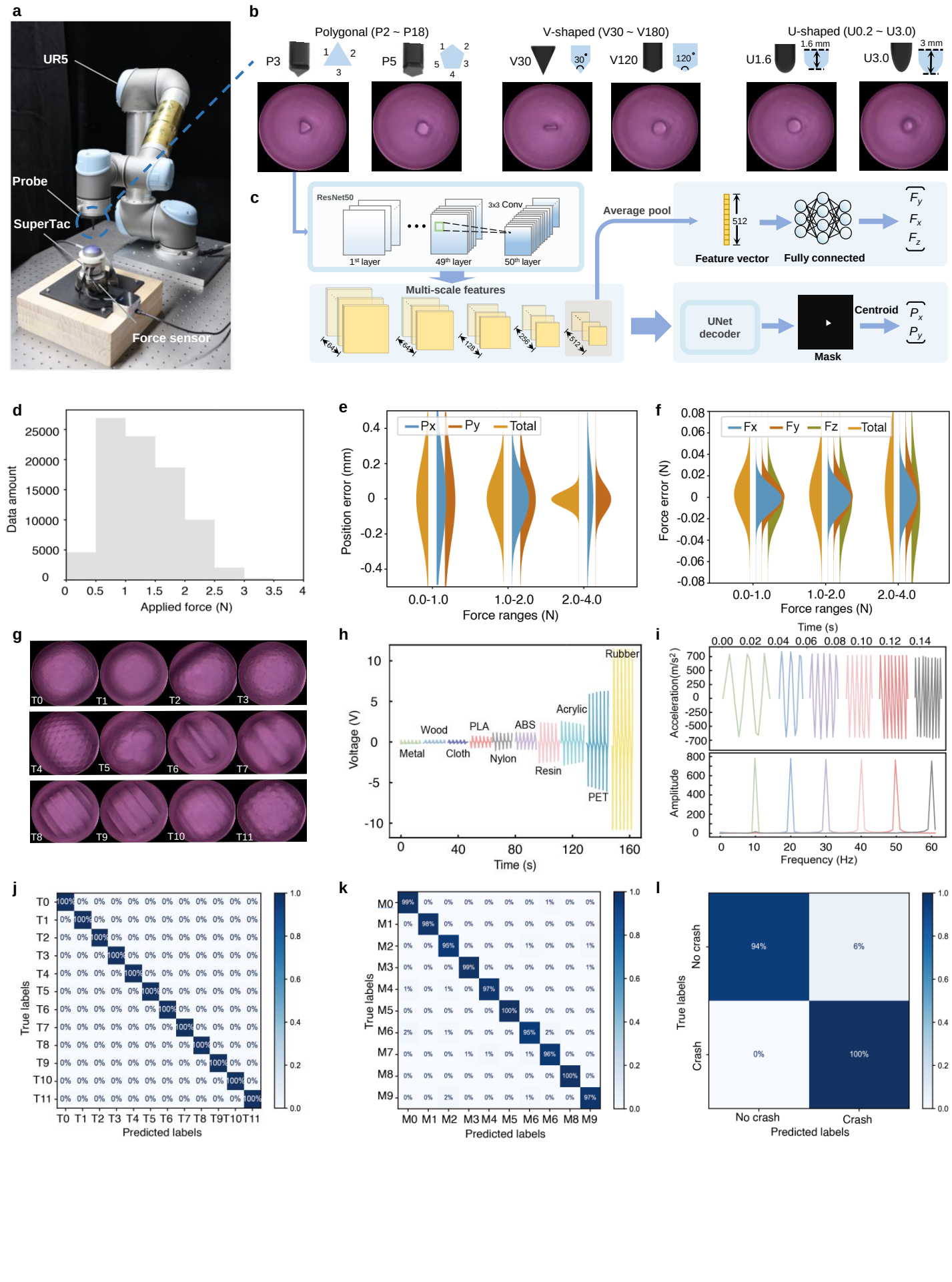
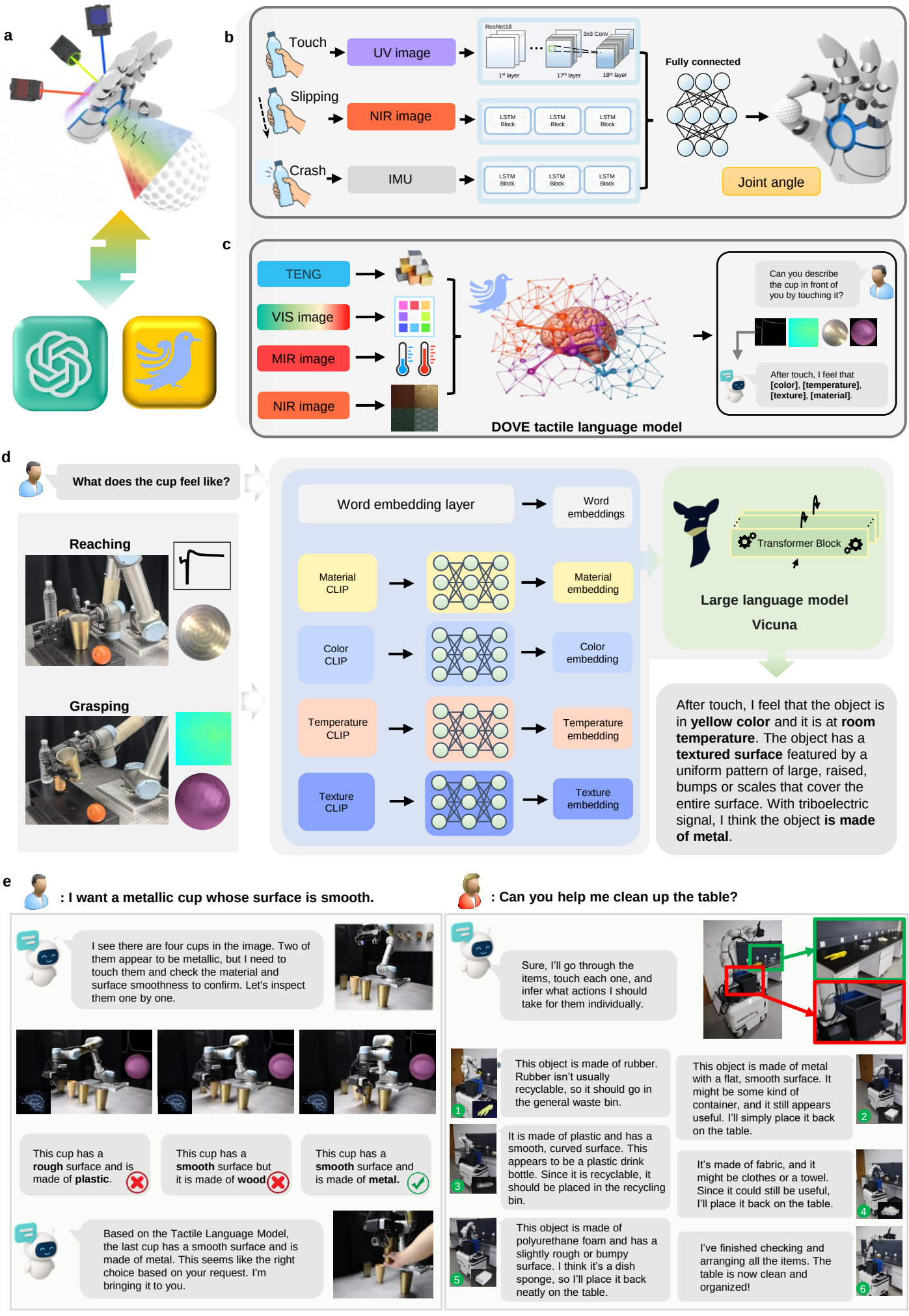


Fig.4 Design and application of tactile language model. (**a**) The integration of SuperTac with DOVE in human-robot interaction. (**b**) Stable object grasping by combining external vision with contact, slide, and collision sensing. (**c**) Fusion of material, texture, color, and temperature information, combined with a tactile language model for tactile information understanding. (**d**) The tactile language model we designed and its application to tactile information understanding. (**e**) Experiments in human-robot interaction utilizing tactile language model. The tactile language model assists robots in decision-making by providing detailed analyses and reasoning of tactile data.


a
Retina
Lens
Light
Cone
Rod
Optic Disk
b
Sensing skin
Multimodal sensing system
Multispectral imaging perception
Retinal photoreceptor cells
330 nm
500 nm
700 nm
c
Multispectral imaging module
350 nm
700 nm
940 nm
5.5 μm
14 μm
Non-imaging perception
Magnetic field receptor
Disc Membrane
Rhodopsin
Cryptochrome
d
Triboelectric sensor
e
Inertial measurement unit
f
Self-posture
Temperature
Color
Proximity
Texture
Material
Vibration
Force
Deformation
Sliding
UV
VIS
NIR
MIR
TENG
IMU
g
Tactile language model (DOVE)
Recognition
Stable grasping
HRI
h
Ref.4
Ref.7
Ref.9
Ref.12
Ref.16
Ref.17
Ref.18
Ref.19
Ref.20
Ref.21
Ref.22
Ref.23
Ref.24
Ref.25
Ref.26
Ref.27
Ref.28
Ref.29
Ref.30
Ref.31
This work
Visuotactile sensor
Single point
Function
Resolution
0.01
0.1
1
10
100
1000


a
1 cm
UV
VIS
NIR
MIR
Triboelectric signal
Pose
Y-axis
X-axis
Z-axis
b
① Conductive layer
② Reflective layer
③ Fluorescent layer
④ Supporting layer
⑤ MIR camera
⑥ NIR LED
⑦ IMU
⑧ PEDOT:PSS
⑨ Fluorescent marker
⑩ Visible camera
⑪ NIR camera
⑫ UV LED
c
Transparent
390 nm
940 nm
UV&Visible
MIR
NIR
Touch mode
Vision mode
d
UV
1 N
NIR
1 cm
VIS
MIR
0°C
90°C
e
Positive charge
Negative charge
TPU
PEDOT: PSS
Silica gel
f
Triboelectric signal
Fitted curve
Contacted force
Error bar
Voltage (V)
Force (N)
Distance (mm)


a
UR5
Probe
SuperTac
Force sensor
b
Polygonal (P2 ~ P18)
P3
P5
V-shaped (V30 ~ V180)
V30
V120
U-shaped (U0.2 ~ U3.0)
U1.6
1.6 mm
U3.0
3 mm
c
ResNet50
3x3 Conv
1st layer
49th layer
50th layer
Multi-scale features
Average pool
512
Feature vector
Fully connected
UNet decoder
Centroid
Mask
d
Data amount
Applied force (N)
e
Px
Py
Total
Position error (mm)
0.0-1.0
1.0-2.0
2.0-4.0
Force ranges (N)
f
Fx
Fy
Fz
Total
Force error (N)
Force ranges (N)
g
T0
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
h
Rubber
Acrylic
Wood
PLA
ABS
Metal
Cloth
Nylon
Resin
PET
Voltage (V)
Time (s)
i
Time (s)
Acceleration(m/s²)
Amplitude
Frequency (Hz)
j
True labels
Predicted labels
k
True labels
Predicted labels
l
No crash
Crash
True labels
Predicted labels


a
b
Touch
UV image
Slipping
NIR image
Crash
IMU
ResNet18
3x3 Conv
1st layer
17th layer
18th layer
LSTM Block
Fully connected
Joint angle
c
TENG
VIS image
MIR image
NIR image
DOVE tactile language model
Can you describe the cup in front of you by touching it?
After touch, I feel that [color], [temperature], [texture], [material].
d
What does the cup feel like?
Reaching
Grasping
Word embedding layer
Word embeddings
Material CLIP
Material embedding
Color CLIP
Color embedding
Temperature CLIP
Temperature embedding
Texture CLIP
Texture embedding
Transformer Block
Large language model
Vicuna
After touch, I feel that the object is in yellow color and it is at room temperature. The object has a textured surface featured by a uniform pattern of large, raised, bumps or scales that cover the entire surface. With triboelectric signal, I think the object is made of metal.
e
: I want a metallic cup whose surface is smooth.
I see there are four cups in the image. Two of them appear to be metallic, but I need to touch them and check the material and surface smoothness to confirm. Let's inspect them one by one.
This cup has a rough surface and is made of plastic.
This cup has a smooth surface but it is made of wood.
This cup has a smooth surface and is made of metal.
Based on the Tactile Language Model, the last cup has a smooth surface and is made of metal. This seems like the right choice based on your request. I'm bringing it to you.
: Can you help me clean up the table?
Sure, I'll go through the items, touch each one, and infer what actions I should take for them individually.
1
This object is made of rubber. Rubber isn't usually recyclable, so it should go in the general waste bin.
2
This object is made of metal with a flat, smooth surface. It might be some kind of container, and it still appears useful. I'll simply place it back on the table.
3
It is made of plastic and has a smooth, curved surface. This appears to be a plastic drink bottle. Since it is recyclable, it should be placed in the recycling bin.
4
It's made of fabric, and it might be clothes or a towel. Since it could still be useful, I'll place it back on the table.
5
This object is made of polyurethane foam and has a slightly rough or bumpy surface. I think it's a dish sponge, so I'll place it back neatly on the table.
6
I've finished checking and arranging all the items. The table is now clean and organized!