Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal tactile–posture capture enables physically grounded annotation of human demonstrations

Lin Fan
Shenzhen University
Herong Wang
Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen)
Yong Tian
Shenzhen University
Yiyang Li
Shenzhen University
Xiaojian Rao
Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen)
Yue Xiao
Sun Yat-sen University
Yuzhou Wang
Sun Yat-sen University
Weiguan Zhang
Sun Yat-sen University
Jindong Tian
Shenzhen University
Ning Li
Sun Yat-sen University
Gaojie Chen
chengj235@mail.sysu.edu.cn
Sun Yat-sen University

Article

Keywords:

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11217490/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multimodal tactile–posture capture enables physically grounded annotation of human demonstrations

Lin Fan[1,7], Herong Wang[2,3,7], Yong Tian[1], Yiyang Li[1], Xiaojian Rao[2], Yue Xiao[3], Yuzhou Wang[5,6], Weiguan Zhang[3,4], Jindong Tian[1,2,✉], Ning Li [3, ✉] and Gaojie Chen[3,✉]

[1]Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, China.

[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen), Shenzhen 518132, China.

[3]School of Flexible Electronics (SoFE) & State Key Laboratory of Optoelectronic Materials and Technologies (OEMT), Sun Yat-sen University, Shenzhen, Guangdong 518107, China.

[4]State Key Laboratory of Flexible Electronics (LoFE), Nanjing University of Posts & Telecommunications, 9 Wenyuan Road, Nanjing 210023, China.

[5]Department of Neurology, the Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou,510655, China.

[6]Biomedical Innovation Center, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou,510655, China.

[7] These authors contributed equally: Lin Fan, Herong Wang.

E-mail: chengj235@mail.sysu.edu.cn

Abstract

Learning dexterous manipulation from human demonstrations requires records of physical interactions that video alone cannot fully capture. Here we present tactile-motion capture (TaMoCap), a portable physical interaction tracking system that records multimodal tactile–posture information during natural human demonstrations, reconstructs physical interaction states of human-object interaction, and converts them into structured semantic annotations. The system integrates a smart textile glove with distributed three-axis force, triboelectric and strain sensors. The synchronized multimodal sensing streams are processed to reconstruct three-axis contact forces, strain-derived joint-angle estimates, and triboelectricity-based contact-transition states, yielding physically interpretable representations that characterize sustained loading, transient interaction events, and hand posture during natural manipulation. A tactile semantic translation framework integrates synchronized visual observations, reconstructed physical interaction states, and task context to generate action annotations, which are verified against multimodal evidence for physical consistency and source traceability. In a descriptive evaluation of four selected massage cases, the full visual–tactile–semantic input achieves a mean composite score of 94.75/100 in model-based assessment, 33.5 points higher than the video-only input. In a long-horizon noodle-cooking demonstration, TaMoCap preserves temporal alignment between multimodal sensing evidence and structured semantic annotations throughout the action sequence. TaMoCap provides a systematic approach for constructing physically grounded human demonstration records for tactile-enhanced embodied agents and robotic manipulation systems.

Introduction

Dexterous manipulation in unstructured environments requires robots to understand task goals and promptly adjust their actions in response to physical feedback, such as changes in contact force and slip[1, 2]. Therefore, demonstration datasets for imitation learning and vision–language–action (VLA) model training must be large-scale and diverse while preserving temporal alignment among sensory observations, actions, and tactile feedback throughout continuous recordings of manipulation[3, 4]. Simulation and robot teleoperation produce action-labelled data, but they remain resource-intensive and may not fully capture the tactile feedback, along with accumulated expertise underlying natural human manipulation[5-7]. First-person recording provides a large-scale egocentric video dataset of everyday human activities, capturing object use, action sequences, and adaptive behavior in a form compatible with vision–language supervision[8-11]. Humans rely on tactile feedback to regulate grip and adjust movements[12-14]. However, video alone provides limited information about contact mechanics, particularly when contact regions are occluded or objects deform[15]. Therefore, first-person demonstration systems should synchronously record video, hand posture, spatially distributed three-axis contact forces, and contact or slip events without restricting natural movement [16-22].

Existing systems for recording human demonstrations primarily focus on hand motion or contact sensing. Portable motion-capture platforms have greatly expanded the range of human activities captured, increasing the diversity of demonstration datasets. However, these recordings remain primarily visual or kinematic, providing limited insight into the underlying contact mechanics. Smart textile gloves combining strain-sensing yarns and inertial measurement units track hand and finger movements [23]. Wearable pressure-sensing gloves can map contact intensity, but pressure-dominant measurements mainly resolve normal loading and do not fully characterize tangential force, frictional state or slip tendency[23-25]. Optical tactile sensors can recover local geometry and detect slip, but they are difficult to distribute across the hand and do not simultaneously measure whole-hand posture or sustained loading during unconstrained daily activity[26-29]. Likewise, event-sensitive slip or contact sensors identify transient

changes but provide limited information about the underlying quasi-static mechanical state[30-34]. Existing systems face trade-offs among tactile sensing coverage across the hand, three-axis contact-force measurement, transient interaction detection, hand-posture tracking, and wearability. An effective recording platform should integrate these complementary sensing capabilities and synchronize the resulting measurements with video while preserving the natural timing of human manipulation.

Multiparameter tactile streams contain rich information about dynamic interactions, including sustained loading, transient contact states and more. Such information is closely related to sensor calibration and location, the interaction object, and the task phase, together with visual observations and hand motion[35]. Raw tactile signals do not directly provide multimodal large language models (MLLMs) with semantic descriptions of such information[36, 37]. Existing approaches encode raw or windowed tactile signals into latent representations for perception and policy learning, sometimes incorporating modality fusion or cross-modal alignment[38-42]. Although useful for prediction, these representations do not provide explicit descriptions of explicit physical semantics, event boundaries or sensor-source provenance[41, 42]. Therefore, enabling MLLMs to understand tactile streams requires reconstructing tactile state and motion trajectories, performing contact-event-driven temporal segmentation, further generating structured annotations that tightly link physical interaction state descriptions to sensing modalities, sensing channels, and task time windows[43]. The resulting annotations support the construction of structured demonstration datasets for subsequent VLA learning.

Herein, we present tactile and motion capture (TaMoCap), an integrated system that that records multimodal human demonstration data, reconstructs physical features of hand–object interactions and translates the sensing evidence into structured semantic annotations. The system records multimodal information from a flexible glove integrating 14 distributed three-axis force sensors, eight triboelectric nanogenerators (TENGs) and 10 dorsal strain sensors, with all streams synchronized to video. The system synchronizes multimodal sensor signals with video and reconstructs time-aligned physical representations of manipulation, including contact force distributions,

triboelectric response states, and hand-posture trajectories. These representations characterize sustained loading, transient interactions, and hand-posture changes.

TaMoCap further incorporates a tactile semantic translation (TaST) framework that combines time-aligned visual, tactile, and postural information with task context to generate action annotations. In a multidimensional assessment of annotation quality across four massage cases, the visual–tactile–semantic input configuration achieved a mean composite score of 94.75/100, 33.5 points above the video-only. In a long-horizon noodle-cooking demonstration, TaMoCap maintained temporal alignment between multimodal sensing evidence and annotations throughout the action sequence. By linking multimodal sensing, reconstructed physical states, and semantic annotations, TaMoCap supports the construction of tactile-enhanced human demonstration datasets for embodied intelligence.

Results

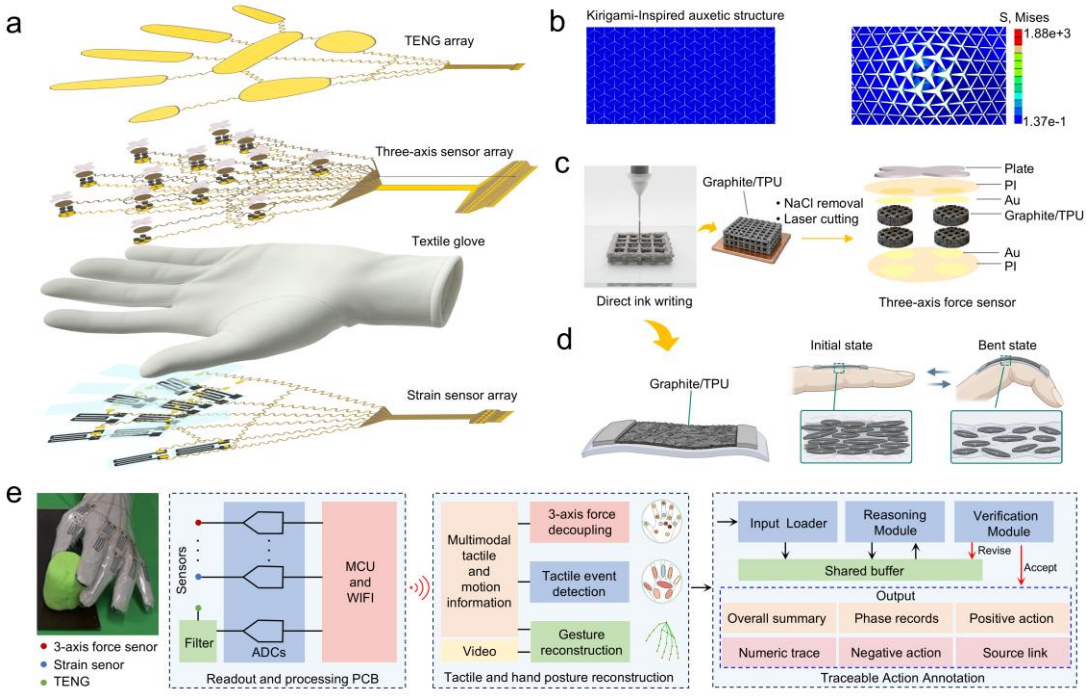


Fig. 1 | Overview of TaMoCap. a Exploded view of the flexible data acquisition glove showing the palmar TENG and three-axis force array, textile glove and dorsal strain array. **b** Kirigami-inspired auxetic TENG geometry and simulated out-of-plane deformation. **c** Direct-ink-writing fabrication of the porous graphite/TPU/NaCl composite and structure illustration of a four-element

three-axis force sensor. **d** Graphite/TPU strain sensor in the initial and bent states, illustrating deformation of the conductive network. **e** TaMoCap workflow from synchronized human demonstration acquisition and wireless transmission to tactile–posture reconstruction, and TaST semantic translation, in which multimodal sensing evidence and task context are integrated through reasoning and verification to generate structured annotations, including recommended actions and actions to avoid.

Overview and design of the TaMoCap

Fig. 1 illustrates TaMoCap, which uses a data glove to capture multimodal tactile and hand-posture information from human demonstrations, and combines it with reconstructed physical interaction states of hand–object interactions to further generates action annotations under the TaST framework. The wearable glove uses a textile substrate to mount the tactile sensing module on the palmar side and the posture sensing module on the dorsal side (Fig. 1a). The sensor arrays connect to wireless readout electronics and are synchronized with an external video stream, enabling the same manipulation segment to be represented by reconstructed three-axis forces, contact-state signals, and hand-motion measurements, as well as visual.

The two sensing layers distributed on opposite sides of the palm include 14 three-axis force sensors and 8 TENGs. A Kirigami-inspired auxetic geometry allows the TENG electrodes to conform to curved surfaces and deform out of plane (Fig. 1b). Each force sensor contains four porous graphite/thermoplastic polyurethane (TPU) elements positioned below a load-transfer plate (Fig. 1c). Their summed and differential responses form the basis of the calibrated force-and-moment-balance model for estimating normal and tangential force under the loading assumptions specified in Supplementary Note 1. Ten graphite/TPU strain sensors located on the dorsal side of the hand use deformation-dependent changes in the conductive network to track finger-joint flexion (Fig. 1d).

The processing and annotation workflow is shown in Fig. 1e. Filtered sensor signals are digitized, packaged and transmitted with video to the reconstruction stage, which incorporates sustained load responses, transient interaction signatures and continuous

hand deformation during natural manipulation. The TaST integrates the synchronized multimodal sensing evidence with task context to generate verified, time-aligned and source-traceable annotations that distinguish sensor-grounded observations from recommended actions and actions to avoid. This workflow retains the temporal correspondence between measured interaction states and their semantic descriptions.

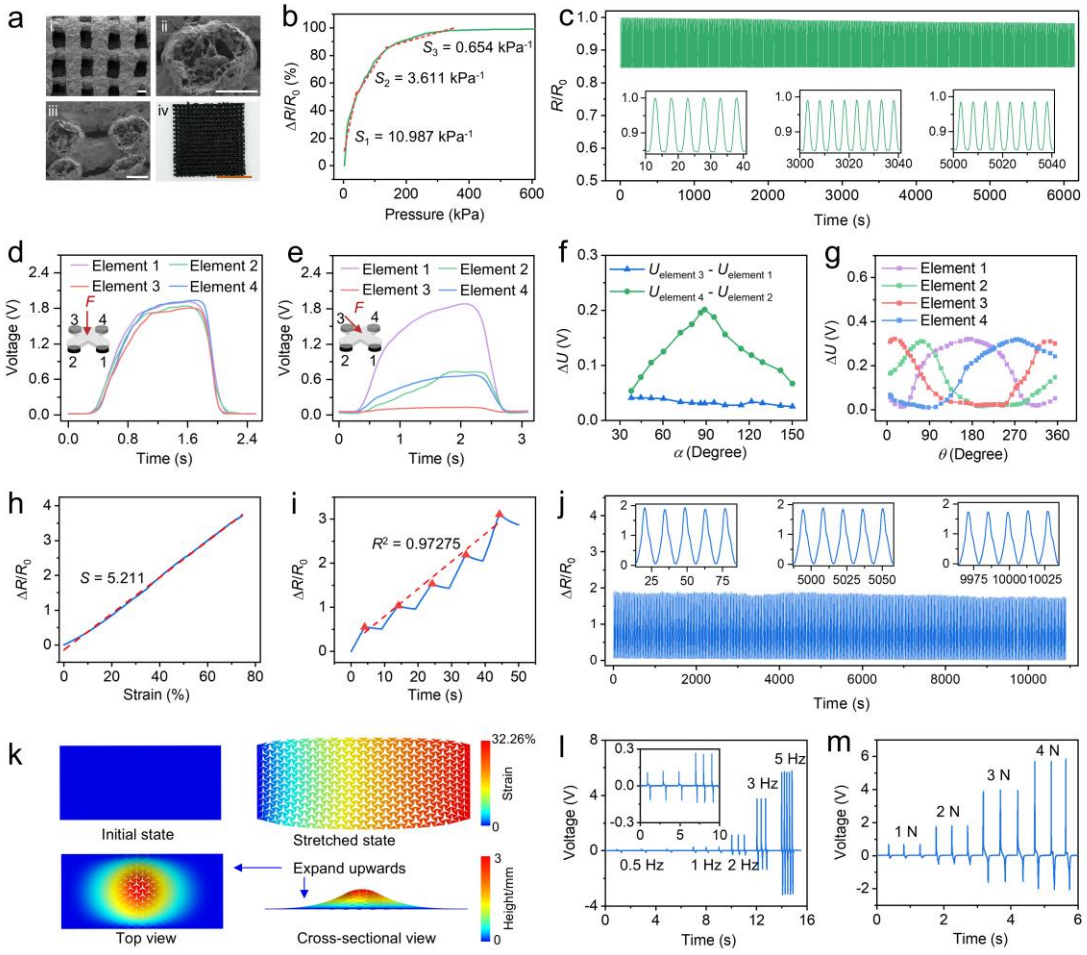


Fig. 2 | Fabrication and performance of multimodal sensors. a Scanning electron microscope images of the hierarchical porous graphite/thermoplastic polyurethane (TPU) composite network structure (i-iii) and photograph of the printed graphite/TPU/NaCl composite (iv), scale bars: 300μm in (i)–(iii) and 1 cm in (iv). **b** Relative resistance changes as a function of pressure, showing sensitivities of 10.987, 3.611 and 0.654 kPa^{-1} in three pressure regimes. **c** Cyclic stability over 6,000 s under an applied pressure of 10 kPa; insets show representative cycles near the beginning, midpoint and end of the test. **d and e** Four element voltage responses under normal force loading (d) and tangential force loading (e). **f** Differential response features of four elements as functions of polar angle α. g Four-element voltage changes over a complete in-plane rotation θ. **h–j** Strain-sensor

response to continuous strain (h), staged deformation (i) and more than 10,000 s test time (j). k Simulated deformation of the Kirigami triboelectric nanogenerator (TENG) electrode in its initial, stretched and upward-expanded states. **l** and **m** TENG voltage responses at contact–separation frequencies of 0.5–5 Hz (l) and dynamic loads of 1–4 N (m).

Sensor fabrication and sensing performance

The pressure-sensitive materials of the three-axis force sensor were fabricated via direct ink writing (DIW) printing (Fig. S1 and Methods), using an ink containing graphite, thermoplastic polyurethane (TPU), dimethylformamide (DMF), and sodium chloride (NaCl) particles. After thermal curing and template removal, a waffle-like hierarchical porous network structure was obtained (Fig. 2a and Fig. S2). The hierarchical porous structure enhances pressure sensitivity through two complementary mechanisms: i) the primary printing-induced pores increase mechanical compliance and facilitate stress transfer, ii) the secondary template-derived micropores amplify local deformation and conductive-network reconstruction, thereby increasing the pressure-induced resistance change (Fig. S3) [44]. As a result, the sensor achieves high sensitivity up to 10.987 kPa^{-1} (Fig. 2b), stable repeatable pressure responses (Fig. 2c and Fig. S4). Through performance optimization, the optimal mass ratio of TPU, graphite and NaCl was determined to be 100%:50%:400%, respectively (Fig. S5).

Three-axis force sensing used the common- and differential-mode responses of four elements arranged beneath a rigid load-transfer structure (Fig. S6a). Under normal loading, the porous network undergoes uniform compression, resulting in consistent common-mode signal variations across the four elements (Fig. 2d and Fig. S6b). Under shear loading, four elements exhibit distinct responses, generating differential-mode responses (Fig. 2e and Fig. S6c). When maintaining a constant loading force while varying the loading angle, the four elements show highly regular differential responses (Figs. 2f and 2g). The output variations of the four sensing elements exhibit a stable correspondence with the applied three-dimensional force, providing a reliable basis for three-axis force decoupling.

The strain sensor achieves high strain sensitivity through controlled microcrack formation within the conductive network during stretching (Fig. 1d and Fig. S7) [45]. Under applied strain, reversible microcrack opening and closing modulate the conductive pathways, thereby inducing the electrical response to deformation. The sensor shows a monotonic resistance change over strains up to approximately 75% (Fig. 2h). Its response follows staged deformation with a coefficient of determination of 0.97275 (Fig. 2i), and the signal amplitude is retained during more than 10,000 s loading and unloading time (Fig. 2j). These measurements support continuous tracking of finger flexion during repeated manipulation.

The multimodal data acquisition glove also incorporates Kirigami-inspired TENGs over the palmar contact regions, in which Polyurethane (PU) film is selected as surface friction layer. Because the Au/Cu electrode is intrinsically non-stretchable, the Kirigami geometry provides structural extensibility and conformability[46]. Finite-element analysis of the Kirigami-structured TENG illustrates how in-plane stretching and localized upward displacement are accommodated by rotation and opening of the auxetic cells (Fig. 2k). The expandable geometric design improves electrode conformability on complex curved surfaces and reduces mechanical constraints during dynamic deformation. Fig. 2l and 2m show that the TENG outputs stable voltage responses under different stimulation conditions. The signal amplitude varies with both stimulation frequency and applied force, demonstrating the its capability to capture transient contact change during dynamic interaction.

The three-axis force, strain, and triboelectric sensor arrays provide multi-parametric information on human–object interactions, including spatially distributed sustained load responses, transient interaction signatures and continuous hand deformation. Their integration provides the multimodal physical inputs required for subsequent physical interaction states reconstruction and semantic annotation.

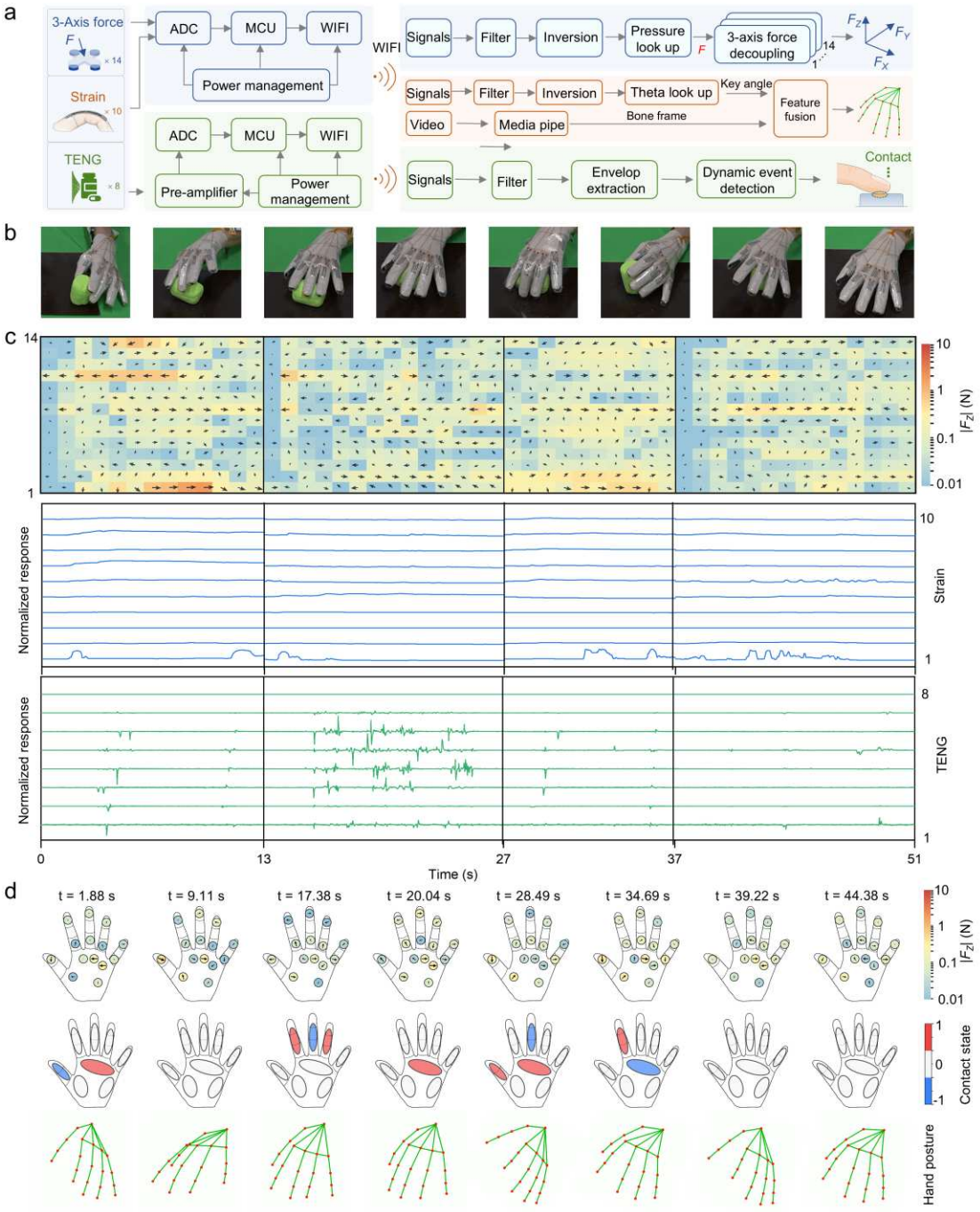


Fig. 3 | Multimodal acquisition and physical interaction state reconstruction during manipulation. a Dual-link acquisition and reconstruction pipeline. Signals from the three-axis force sensors and strain sensors are digitized and transmitted over one link different from that used for the triboelectric signals; the synchronized signals are represented as reconstructed three-axis forces and hand posture, and transient interaction signatures. **b** Video frames from consecutive operations of pinching, rolling, flipping and pressing performed on a modelling clay. **c** Reconstructed three-axis force maps (top), normalized strain responses (middle) and normalized triboelectric responses (bottom) across the time sequence; color intensity represents the magnitude of the reconstructed

sensor-normal force and arrows represent the two reconstructed in-plane force components; vertical lines mark action boundaries. **d** Reconstructed palmar three-axis force maps, TENG response maps and reconstructed hand posture at selected time points; 1 denotes transient contact, 0 denotes no change in the contact state, and −1 denotes transient separation.

Glove integration and reconstruction of physics interaction states

The multi-parametric data acquisition glove integrates 14 three-axis force sensors and 8 TENGs at the principal palmar contact regions, and 10 strain sensors at dorsal finger joints (Fig. S8). Because the output of individual channels in the TENG array is susceptible to capacitive coupling, the signal lines are shielded. Sequential activation of individual channels produces negligible response in inactive channels, indicating suppression of inter-channel crosstalk (Fig. S9). Distinct strain response patterns are also obtained when a participant makes gestures such as “OK” and “yeah” (Fig. S10), confirming that the responses of strain sensor array distinguish different gestures. The synchronized sensor signals are reconstructed into three-axis forces, hand posture and transient interaction signatures by the proposed acquisition and reconstruction strategies (Fig. 3a).

Contact force reconstruction begins with the four baseline-subtracted voltages (Eq. 1). A predefined nonlinear response function maps each voltage change to an element-force estimate (Eq. 2). These estimates are then combined through opposing-element differences and the four-element sum to reconstruct the three-axis force based on force and moment balance (Eq. 3). The eight element-response coefficients ($a_i, c_i, i = 1,2,3,4$) and the differential-response calibration coefficient (η) are jointly determined by equal-weight least-squares fitting to reference measurements (Supplementary Note 1). Fig. S11 compares the reconstructed force components with the reference over the calibration interval, providing an assessment of the model fit. Further validation across independent loading cycles and sensing units would help assess its broader applicability.

$$U = [U_1, U_2, U_3, U_4]^T,\ U_i = U_{\text{meas},i} - U_{\text{base},i},\ i = 1,2,3,4 \qquad \text{Eq. 1}$$

$$f_i(U_i) = \frac{a_i U_i}{1 - c_i U_i}, i = 1,2,3,4 \qquad \text{Eq. 2}$$

$$\hat{\boldsymbol{F}} = \boldsymbol{G}\boldsymbol{f}\left(\boldsymbol{U}\right), \boldsymbol{G} = \begin{bmatrix} \eta & 0 & -\eta & 0 \\ 0 & \eta & 0 & -\eta \\ -1 & -1 & -1 & -1 \end{bmatrix} \quad \text{Eq. 3}$$

Here $V_{i,0}$ is the median voltage over the initial 0.5 s of the calibration interval, a_i and c_i have units of N/V and V^{-1}, respectively, and η is dimensionless. The vector $f\left(U\right)$ contains the four element forces, and $\hat{F} = \left[\hat{F}_x, \hat{F}_y, \hat{F}_z\right]^T$ gives device-frame force components, with external compression corresponding to negative F_z. Eq. 3 assumes slow, center loading with negligible inertia and no independent contact couple; eccentric normal loading can affect the tangential-force estimates (Supplementary Note 1, Section 1.3).

Meanwhile, an arc-based strain-to-joint conversion model (Supplementary Note 2) converts strain responses into joint-angle (θ_j) estimates based on the calibrated relationship between mechanical strain and electrical signals, as shown in Eq. 4,

$$\theta_j(\mathrm{t})\left[^\circ\right] = \frac{180}{\pi}\frac{L_{j,0}}{r_j}\varepsilon_j(t) \quad \text{Eq. 4}$$

where j indexes the instrumented joint, t is time, $L_{j,0}$ is the effective dorsal coverage length in the worn reference posture, r_j is the effective bending radius, and ε_j is the fractional change in covered length relative to that posture. Equation 4 assumes an extended, zero-angle reference and constant effective radius; the worn-reference strain definition and voltage – strain conversion are specified in Supplementary Note 2, Sections 2.1 and 2.2. The procedure for incorporating these estimates into a visual hand kinematic model is specified in Supplementary Note 2, Section 2.3 and Fig. S12. It remains to be further studied whether the present posture reconstructions can establish absolute joint-angle accuracy.

TENG signals are encoded into positive, within-threshold and negative response states by threshold-based classification (Supplementary Note 3, Section 3.6) and interpreted together with the synchronized video.

We evaluated the reconstruction results during four consecutive operations performed on a clay model: pinching, rolling, flipping and pressing (Fig. 3b). These actions contain localized loading, continuous posture adjustment and repeated contact transitions. Fig. 3c shows the synchronized multimodal sensing signals. The

reconstructed force maps capture changes in loading and its spatial distribution in sensor coordinates. Triboelectric transients accompany the demonstrated manipulations, whereas the strain channels track action-dependent finger deformation. Selected time points are further rendered as palmar reconstructed three-axis force maps, TENG response maps and reconstructed hand skeletons (Fig. 3d). The streams demonstrate how the combination of acquisition and reconstruction in TaMoCap separates sustained loading, transient interaction signatures, and hand configuration within a natural manipulation sequence.

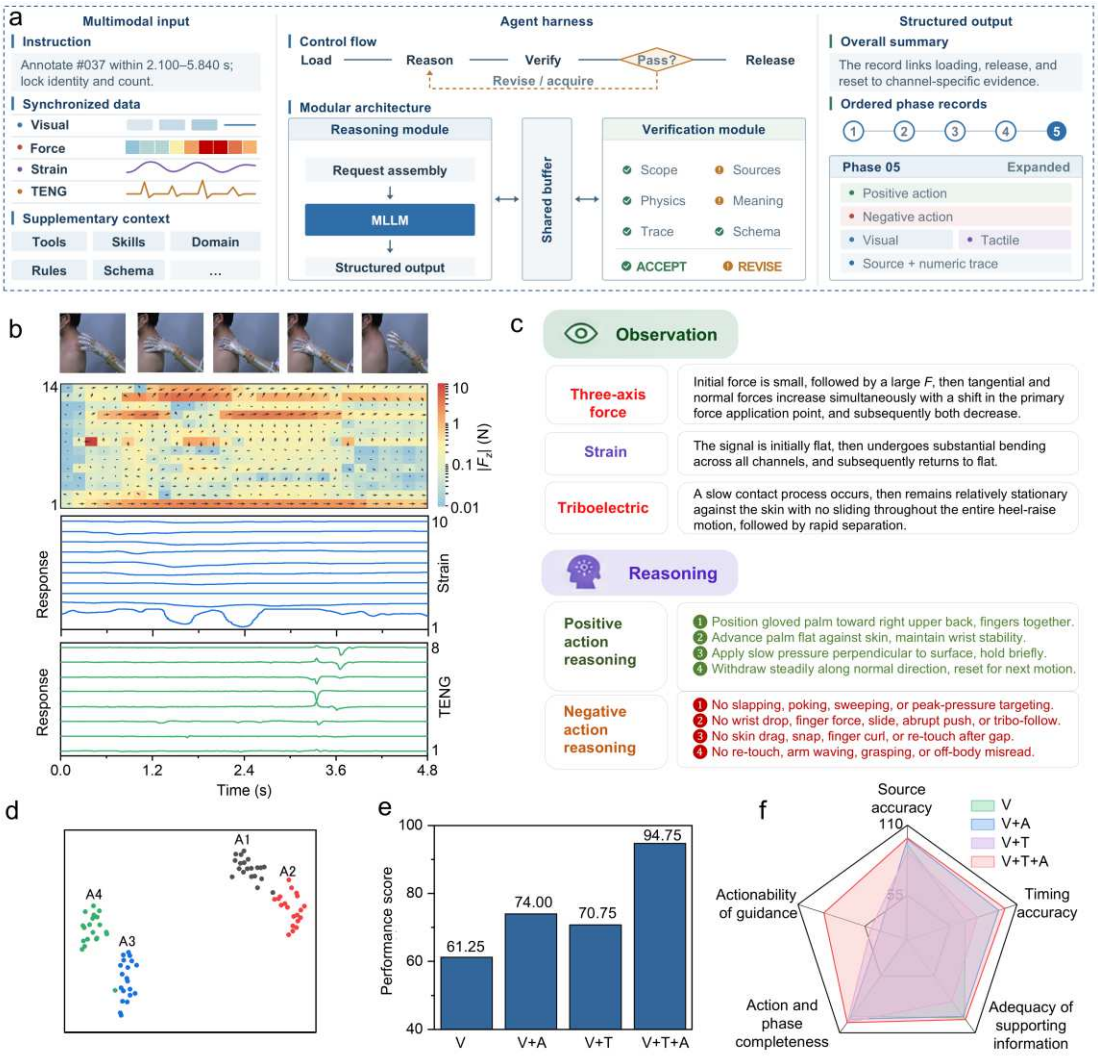


Fig. 4 | TaST framework and evaluation of structured action annotations. a TaST receives an instruction, synchronized visual, reconstructed three-axis force, strain and TENG signals, and supplementary task context; its reasoning and verification modules iteratively generate and check source-linked candidates before releasing an overall summary and ordered phase records. **b** Video

frames and synchronized reconstructed force, strain and triboelectric signals from a 4.8-s massage segment. **c** Example output separating sensor-grounded observations from recommended actions and actions to avoid. Event-specific interpretations in the model output are not validated by the TENG threshold encoding. **d** t-SNE visualization of classification results for four massage actions based on multimodal sensing signals. **e** Composite annotation scores obtained under four input conditions: video only (V), video and structured annotations (V+A), video and tactile (V+T), and video, tactile and structured annotations (V+T+A). Stacked bars represent citation accuracy and annotation guidance value (50 points each). **f** Five-axis radar plots showing source accuracy, timing accuracy, adequacy of supporting information, action and phase completeness, and actionability of guidance. Scores are averaged across the four cases and expressed as percentages of the maximum score for each axis. Panel (d) uses 80 segments (20 per action class). Panels (e and f) report means across four manually selected cases, one per action class, with one model-based score per candidate.

Time-aligned structured semantic annotations for actions

TaST uses synchronized visual and tactile evidence to generate temporally grounded, verifiable action annotations (Fig. 4a). The reasoning module integrates video, reconstructed three-axis forces, strain-derived hand posture, triboelectric response and task constraints to generate a candidate annotation in a predefined structure, with explicit links to its supporting evidence. The verification module evaluates task scope, source availability, physical consistency, semantic validity, source traceability and schema compliance. Candidates that fail any check are returned for revision or acquisition of additional evidence.

For each record, the instruction specifies the action to be annotated and the accompanying, The evidence package pairs time-stamped video with sustained loading, transient interaction signatures, and hand posture, all three synchronized with one other. Supplementary context provides the relevant tools, skills, domain knowledge, task rules and output schema, thereby defining both the evidential scope and the required structure of the annotation.

The reasoning module combines the instruction, evidence, rules and schema into a single MLLM request and generates a candidate annotation, while a shared buffer retains the control state, intermediate records and iteration history. Candidates that pass verification are released as source-linked records comprising an overall summary, an

ordered sequence of action phases and an expanded record for each phase. Each phase stores recommended actions, actions to avoid, visual and tactile observations, and the sources supporting those observations in separate fields. This structure distinguishes direct observations and sensor responses from rule-based guidance, while preserving the physical and temporal basis of each statement.

In a representative 4.8 s segment of massage performed on the participant (Fig. 4b), TaST places the visual and tactile information on a common timeline and links each description to the corresponding signal segment in each modality. Based on the observed behavior and task constraints, it then identifies recommended actions and actions to avoid for subsequent robot execution (Fig. 4c). The massage dataset comprises four action classes with 20 repetitions per class (n = 80 segments, Fig. S13). A t-SNE projection of multichannel time series from the 80 segments provides a descriptive visualization of the four action classes (Fig. 4d) [47]; labels are used only for coloring. We evaluated four input conditions: video only (V), video and structured annotation (V+A), video and tactile (V+T), and video, tactile and structured annotation (V+T+A). The corresponding composite scores are 61.25, 74.00, 70.75 and 94.75 (Fig. 4e). Relative to V, V+T+A scores 33.5 points higher, comprising differences of 4.75 points in citation accuracy and 28.75 points in annotation guidance value. Relative to V+T, the difference is 24.0 points, comprising 7.75 points in citation accuracy and 16.25 points in annotation guidance value. These differences describe four manually selected cases under the recorded input configurations; the tactile direct and framework results used different initial drafts and input descriptions (Supplementary Note 4, Section 4.5). Within these evaluated cases and configurations, V+T+A input achieved the highest score under the specified rubric. Fig. 4f resolves the composite scores into five sub-score groups, showing that the conditions differed mainly in actionability of guidance.

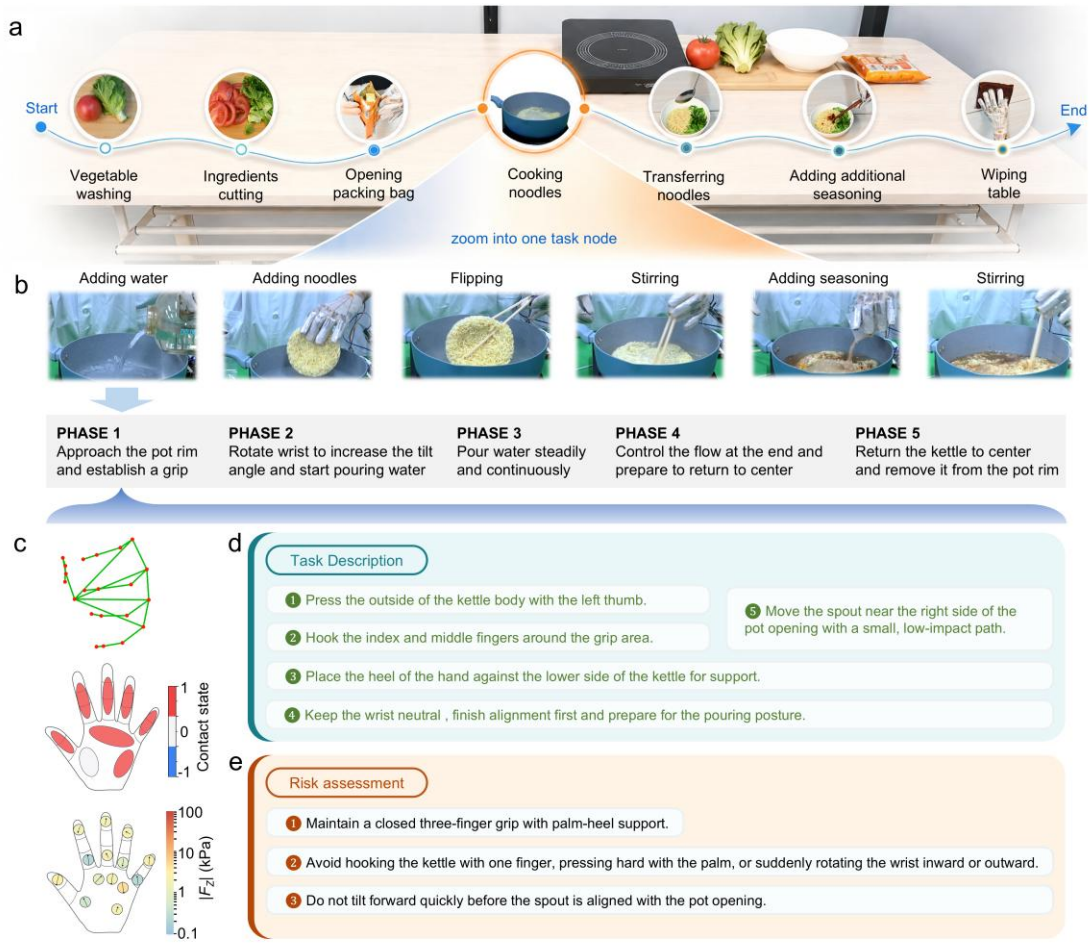


Fig. 5 | Multimodal physical–semantic demonstration records for a long-horizon kitchen demonstration. a Kitchen-workflow timeline providing context for the noodle-cooking annotation example; the background shows the constructed kitchen environment. **b** Video frames of six sequential actions within the cook-noodle task and time-aligned annotations for five execution phases of the add-water action. **c–e** Physical–semantic records for the phase of adding water, including reconstructed hand posture, triboelectric response-state map and reconstructed three-axis force map (c), task description (d), and handling precautions (e).

Multimodal physical–semantic demonstration records for a long-horizon demonstration

Long-horizon demonstrations require local action descriptions to remain connected to their context and physical evidence. We recorded a continuous kitchen sequence comprising vegetable washing, ingredient cutting, opening the noodle packet, cooking noodles, transferring noodles, adding additional seasoning and wiping the table (Fig. 5a). Taking noodle cooking as the annotation example, the selected interval contains

six ordered actions: adding water, adding noodles, flipping the noodles, stirring the noodles, adding seasoning and stirring again (Fig. 5b and Fig. S14), and time-aligned structured annotations are generated for the execution phases of each action. All annotations are anchored to a single timeline keeps each action temporally bounded yet still embedded in the continuous sequence. Movie S2 present the complete noodle-cooking demonstration and allow the structured annotations to be inspected for each segment and time point.

The add-water action is segmented into five phases (Fig. 5b). The task descriptions of phase 1 associate the hand posture and active contact regions (Fig. 5c) with descriptions of thumb contact, index–middle-finger engagement, palm-heel support, neutral wrist posture and spout–pot alignment (Fig. 5d). Rule-based handling precautions are presented separately, including the need for a closed three-finger grasp and the avoidance of single-finger hooking, excessive palmar force, abrupt wrist rotation and tilting before alignment.

Throughout the noodle-cooking interval, each annotation remains linked to time-localized physical evidence. The resulting record retains contact location, reconstructed force state, hand configuration, object relations and operational constraints rather than only an action label. This time-aligned organization supports interpretation of complex, long-horizon operations and provides a basis for fine-grained, contact-grounded demonstration datasets.

Discussion

TaMoCap integrates wearable sensing, physical interaction state reconstruction and semantic annotation in an end-to-end workflow for dataset generation. Its wearable acquisition component combines spatially distributed three-axis force sensors, TENG sensors and dorsal strain sensors with synchronized video. Each modality has a distinct role: the force array records sustained loading through summed and differential responses, the TENG array encodes positive and negative transient responses, and strain sensing provides posture-related measurements alongside vision. The acquisition-and-reconstruction pipeline provides readouts of sustained loading, transient interaction signatures, and hand configuration during natural manipulation sequences. TaST then operates on these reconstructed physical interaction states together with video and task context to produce verified, source-traceable structured annotations.

In the massage evaluation, the complete visual–tactile–semantic condition scored 94.75 versus 61.25 for video alone, a descriptive difference of 33.5 points in four selected cases under the composite rubric. In the noodle-cooking example, TaMoCap maintained temporal alignment among video, tactile and posture states, and phase-level annotations throughout the selected interval. These results show that the system converts natural human demonstrations into structured, contact-grounded data that can support tactile-aware model training, action understanding and the evaluation of embodied agents in extended tasks.

Further validation is required before large-scale deployment. Sensor calibration should be assessed across wearers and recording sessions, and the effects of material type, environmental conditions and visual occlusion on reconstructed states need to be quantified. TaST should be evaluated in repeated blinded studies, followed by closed-loop experiments that test whether TaMoCap data improve robot policy learning and task execution. Further miniaturization of the acquisition electronics and improved wireless stability will facilitate broader data collection.

Methods

Materials

Graphite powder (500 mesh, 99 wt%) was purchased from Macklin Technology Co., Ltd. Thermoplastic polyurethane (TPU) pellets (Wanthane WHT-1185) were obtained from Wanhua Chemical Group Co., Ltd. (Yantai, China). Sodium chloride (NaCl, AR) was supplied by Sinopharm Chemical Reagent Co., Ltd. Flexible printed circuit (FPC) substrates and copper-clad laminates were procured from Shenzhen Jinpeng Electronics. All reagents were used as received without further purification.

Fabrication of the sensor arrays

The pressure-sensitive units were fabricated via direct ink writing (DIW) 3D printing. A composite ink was prepared with a graphite:TPU:NaCl mass ratio of 50:100:400, excluding the solvent. TPU was dissolved in N,N-dimethylformamide at a TPU: solvent mass ratio of 1:4. Graphite powder was dispersed in the solution, and NaCl particles (average size ~50 μm) were added as the sacrificial template and homogenized by mechanical stirring. The ink was loaded into a syringe and printed through a 340 μm nozzle onto a flexible substrate with four Au/Cu electrodes. The printed structures were thermally cured at 80°C for 1 h, then immersed in deionized water at 80°C for 12 h to leach out the NaCl template, forming a hierarchical porous network structure. A force-conducting structure was assembled on top of four porous graphite/TPU elements.

Strain sensors were fabricated by DIW printing using a graphite/TPU composite ink (graphite mass fraction 40%). The ink was printed as linear serpentine patterns on a TPU film. After curing at 80 ˚C for 1 h, the sensors were pre-stretched to 5% strain to initiate the controlled microcrack network in the conductive pathways, which significantly enhanced the gauge factor.

The TENGs were fabricated by laser-cutting Au/Cu electrodes on a FPC substrate to form a kirigami-inspired auxetic structure, with a polyurethane (PU) film serving as the surface triboelectric layer. Additional grounded layers above and below each sensing channel formed a surrounding shielding structure to suppress inter-channel crosstalk arising from parasitic capacitive coupling.

Characterization and testing

Morphologies of the porous structures and microcracks were characterized by scanning electron microscopy (SEM, SU8020, Hitachi). Mechanical tensile and compression tests were conducted on a universal testing machine (ZQ-950B, Zhiqu Precision). The pressure-sensing performance was evaluated using a compression testing machine combined with a multichannel multimeter (Keithley DAQ6510). For three-axis force characterization, a six-axis force/torque sensor (ATI Nano17) was used as a reference; the sensor was subjected to normal and tangential loads at various angles, and the four-channel differential/common-mode responses were recorded. Strain sensors were calibrated on a linear motor stage with a high-precision displacement gauge. TENG sensors were tested under dynamic contact-separation and sliding motions using a shaker (exciter) at controlled frequencies and amplitudes. All measurements were performed at room temperature (25 °C) and relative humidity (40–50%). Durability was evaluated under repeated loading and unloading for 6,000 s for the pressure sensor (Fig. 2c) and more than 10,000 s for the strain sensor (Fig. 2j).

Glove assembly and system integration

All three kinds of sensing arrays were integrated onto a fabric glove (nylon-spandex blend). The three-axis force sensors (14 units) and TENG sensors (8 units) were strategically placed on the palmar side covering fingertips, finger pads, bases, and palm. The strain sensors (10 units) were arranged on the dorsal side at the metacarpophalangeal and interphalangeal joints of the thumb, and the metacarpophalangeal and proximal interphalangeal joints of the other four fingers. Interconnections were routed via flexible flat cables to a customized dual-link data acquisition board. The voltage signals from the 3D force and strain sensors were recorded by a 128-channel acquisition card with a nominal sampling rate of 1 kHz and transferred to the computer through WLAN. The TENG signals were conditioned by charge amplifiers (bandwidth 0.1–10 kHz) and digitized by a high-speed ADC and transferred through WLAN. Video was recorded from an external camera. The camera and both acquisition devices were started together, and their recording origins defined the temporal correspondence.

Signal-feature Visualization

The recorded sensor mapping selected 56 pressure-sensing channels from 14 force units, ten strain channels and eight TENG channels. Each channel was linearly interpolated to 256 uniformly spaced relative-time points spanning its segment. Each channel was standardized using a common mean and population standard deviation across all 80 segments and all relative-time points. The 74 channel waveforms were concatenated in a fixed channel and temporal order, giving 18,944 input values per segment. Two-dimensional t-SNE used Euclidean distance, perplexity 12, PCA initialization, learning rate 50, early exaggeration 12, 1000 iterations, random seed 42 and the Barnes–Hut method with angle 0.5 (scikit-learn 1.9.0). Action labels were used only for coloring; no class-dependent coordinate adjustment was applied. The plot describes waveform similarity over relative action time; absolute duration was not included in the projection.

Massage annotation evaluation

The reported massage annotation and model-scoring runs used an MLLM with reasoning capabilities. For massage evaluation, one representative segment per action class was selected manually; the four selected segments were compared under four conditions: video only (V), video with the semantic condition (V+A), video plus tactile signals (V+T), and video plus tactile signals with the semantic condition (V+T+A). Each candidate received one model-based assessment using the eleven rubric criteria in Supplementary Note 4, Section 4.4. For reporting, the recorded sub-scores were aggregated into citation accuracy (50 points) and annotation guidance value (50 points). This aggregation preserved the original criterion-level scores and composite totals without regenerating or rescoring any candidate, using the rubric in Supplementary Note 3. These are descriptive comparisons of four cases. No inferential hypothesis test was applied. Additional deformable-object manipulation and cooking tasks were used to examine continuous multimodal recording and long-horizon phase annotation.

Tactile semantic translation framework

For each action, the video sequence was uniformly resampled to 30 time-stamped frames, and the massage evaluation embedded 100-Hz series of 42 pressure-response components from the archived input tables and eight raw and baseline-centred TENG channels. The archived evaluation inputs use kPa-labelled common and differential pressure responses; they are distinct from the calibrated Newton-valued force maps described in Supplementary Note 1. The framework could additionally read the full source tables, which included strain channels; the direct prompt did not embed these channels. Each TaST session received (i) an instruction defining the record, time window, action label and expected number of events; (ii) synchronized video and, where specified, the archived pressure-response, strain and TENG data; and (iii) supplementary context comprising tools, skills, domain knowledge, task rules and the JSON schema. The input loader assembled the context, and the reasoning module formulated an MLLM request. Each action was divided into three to five stages; each stage contained an identifier, start and end times, a name, executable task description, actions to avoid, visual observations, tactile measurements and confidence. For tactile evidence, each numerical citation specified the data source, timestamp, sensor channel, value and interpretation. In framework runs, candidate annotations were written to a shared buffer. The shared buffer retained the control state, intermediate records and iteration history. The verification module checked task scope, source availability, physical consistency, semantic validity, source traceability and schema compliance. Candidates that failed any check were returned to the reasoning module for revision or retrieval of additional evidence; candidates that passed were released as final annotations.

Data availability

The data supporting the figures, reconstructions, annotations and evaluation scores reported in this study are available from the corresponding authors upon reasonable request.

Code availability

The code used for data processing, reconstruction, annotation, evaluation and figure generation is available from the corresponding authors upon reasonable request.

Supplementary information

Supplementary information is available with this article.

Movie S1: Tactile visualization during modelling-clay manipulation.

Movie S2: TaMoCap-generated multimodal annotations for a long-horizon task: a noodle-cooking example.

Author contributions

L.F. and H.W. contributed equally to this work. L.F. and H.W. conceived the study and designed the experiments. L.F. and H.W. fabricated the three-axis force sensors, strain sensors and Kirigami TENGs. L.F. organized the overall framework of the study, revised the manuscript, and took the lead in the sensor fabrication and glove hardware. H.W. developed the TaST framework, performed the massage annotation evaluation and analyzed the data. Y.L. and X.R. developed the physical interaction state reconstruction, including the three-axis force reconstruction model, the strain-to-joint-angle conversion and the TENG response-state encoding. Y.X. performed the finite-element simulations of the Kirigami TENG electrode (Figs. 1b and 2k). Y.W. and N.L. designed the massage tasks and the human experiments, studied the mechanisms of human tactile perception, and analyzed the experimental data. G.C. guided the experiments, implemented the experimental scenarios and led the overall system design. L.F. and N.L. wrote the manuscript. Y.T. and W.Z. critically reviewed and revised the manuscript. J.T. and N.L. supervised the research, directed the project and acquired funding. All authors reviewed and approved the final manuscript.

Acknowledgements

The authors sincerely acknowledge financial support from Open research fund of State Key Laboratory of Flexible Electronics, Joint Research Fund of the Sixth Affiliated Hospital and School of Flexible Electronics, Sun Yat-sen University (2026JRF6H-FE003), Guangdong Basic and Applied Basic Research Foundation (No.

2024A1515010164 and 2025A1515010204), and National Natural Science Foundation of China (NSFC Grant No. 62075140 and 62374111). All studies on human subjects were conducted in accordance with ethical guidelines following approval granted by the Medical Ethics Committee of the Medical Department of Shenzhen University (PN-202600177). The informed consent was obtained from each subject.

Competing interests

The authors declare no competing interests.

References

1. Ravichandar, H., Polydoros, A. S., Chernova, S., Billard, A. Recent advances in robot learning from demonstration. *Annu. Rev. Control. Robotics Auton. Syst.* **3**, 297-330 (2020).
2. Kroemer, O., Niekum, S., Konidaris, G. A review of robot learning for manipulation: Challenges, representations, and algorithms. *J. Mach. Learn. Res.* **22**, 1-82 (2021).
3. Triantafyllidis, E., Acero, F., Liu, Z., Li, Z. Hybrid hierarchical learning for solving complex sequential tasks using the robotic manipulation network roman. *Nat. Mach. Intell.* **5**, 991-1005 (2023).
4. Shao, L., Migimatsu, T., Zhang, Q., Yang, K., Bohg, J. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *Int. J. Robot. Res.* **40**, 1419-1434 (2021).
5. Ju, H., Juan, R., Gomez, R., Nakamura, K., Li, G. Transferring policy of deep reinforcement learning from simulation to reality for robotics. *Nat. Mach. Intell.* **4**, 1077-1087 (2022).
6. Bharadhwaj, H.*, et al.* Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024: IEEE; 2024. p. 4788-4795.
7. Si, W., Wang, N., Yang, C. A review on manipulation skill acquisition through teleoperation-based learning from demonstration. *Cognitive Computation and Systems* **3**, 1-16 (2021).
8. Damen, D.*, et al.* The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 4125-4141 (2020).
9. Núñez-Marcos, A., Azkune, G., Arganda-Carreras, I. Egocentric vision-based action recognition: A survey. *Neurocomputing* **472**, 175-197 (2022).
10. Rodin, I., Furnari, A., Mavroeidis, D., Farinella, G. M. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding* **211**, 103252 (2021).
11. Grauman, K.*, et al.* Ego4d: Around the world in 3,000 hours of egocentric video. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022; 2022. p. 18995-19012.

12. Sobinov, A. R., Bensmaia, S. J. The neural mechanisms of manual dexterity. *Nat. Rev. Neurosci.* **22**, 741-757 (2021).
13. Handler, A., Ginty, D. D. The mechanosensory neurons of touch and their mechanisms of activation. *Nat. Rev. Neurosci.* **22**, 521-537 (2021).
14. Gueorguiev, D., Lambert, J., Thonnard, J.-L., Kuchenbecker, K. J. Normal and tangential forces combine to convey contact pressure during dynamic tactile stimulation. *Sci. Rep.* **12**, 8215 (2022).
15. Delhaye, B. P.*, et al.* High-resolution imaging of skin deformation shows that afferents from human fingertips signal slip onset. *Elife* **10**, e64679 (2021).
16. Willemet, L., Huloux, N., Wiertlewski, M. Efficient tactile encoding of object slippage. *Sci. Rep.* **12**, 13192 (2022).
17. Wei, Y.*, et al.* Human tactile sensing and sensorimotor mechanism: from afferent tactile signals to efferent motor control. *Nat. Commun.* **15**, 6857 (2024).
18. Zhou, Z.*, et al.* Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays. *Nat. Electron.* **3**, 571-578 (2020).
19. Li, G., Liu, S., Wang, L., Zhu, R. Skin-inspired quadruple tactile sensors integrated on a robot hand enable object recognition. *Sci. Robot.* **5**, eabc8134 (2020).
20. Wang, G.*, et al.* Printed sensing human-machine interface with individualized adaptive machine learning. *Sci. Adv.* **11**, eadw3725 (2025).
21. Zhao, J.*, et al.* Passive and space-discriminative ionic sensors based on durable nanocomposite electrodes toward sign language recognition. *ACS Nano* **11**, 8590-8599 (2017).
22. Wu, X.*, et al.* Innervate commercial fabrics with spirally-layered iontronic fibrous sensors toward dual-functional smart garments. *Adv. Sci.* **11**, 2402767 (2024).
23. Tashakori, A.*, et al.* Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves. *Nat. Mach. Intell.* **6**, 106-118 (2024).
24. Jiang, C.*, et al.* Capturing forceful interaction with deformable objects using a deep learning-powered stretchable tactile array. *Nat. Commun.* **15**, 9513 (2024).
25. Sundaram, S.*, et al.* Learning the signatures of the human grasp using a scalable tactile glove. *Nature* **569**, 698-702 (2019).
26. Sun, H., Kuchenbecker, K. J., Martius, G. A soft thumb-sized vision-based sensor with accurate all-round force perception. *Nat. Mach. Intell.* **4**, 135-145 (2022).
27. Son, C.*, et al.* Behavioral biometric optical tactile sensor for instantaneous decoupling of dynamic touch signals in real time. *Nat. Commun.* **15**, 8003 (2024).
28. Lambeta, M.*, et al.* Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robot. Autom. Let.* **5**, 3838-3845 (2020).
29. Sui, R., Zhang, L., Li, T., Jiang, Y. Incipient slip detection method with vision-based tactile sensor based on distribution force and deformation. *IEEE Sens. J.* **21**, 25973-25985 (2021).
30. Bai, N.*, et al.* A robotic sensory system with high spatiotemporal resolution for texture recognition. *Nat. Commun.* **14**, 7121 (2023).
31. Xu, G.*, et al.* Self-powered electrotactile textile haptic glove for enhanced human-

machine interface. *Sci. Adv.* **11**, eadt0318 (2025).
32. Xie, X.*, et al.* Neuromorphic computing-assisted triboelectric capacitive-coupled tactile sensor array for wireless mixed reality interaction. *ACS Nano* **18**, 17041 (2024).
33. Yiu, C.*, et al.* Skin-interfaced multimodal sensing and tactile feedback system as enhanced human-machine interface for closed-loop drone control. *Sci. Adv.* **11**, eadt6041 (2025).
34. Zhu, M.*, et al.* Haptic-feedback smart glove as a creative human-machine interface (HMI) for virtual/augmented reality applications. *Sci. Adv.* **6**, eaaz8693 (2020).
35. Mao, Q., Liao, Z., Yuan, J., Zhu, R. Multimodal tactile sensing fused with vision for dexterous robotic housekeeping. *Nat. Commun.* **15**, 6871 (2024).
36. Lee, M. A.*, et al.* Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. 2019 International conference on robotics and automation (ICRA); 2019: IEEE; 2019. p. 8943-8950.
37. Mon-Williams, R., Li, G., Long, R., Du, W., Lucas, C. G. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nat. Mach. Intell.* **7**, 592-601 (2025).
38. Yang, F.*, et al.* Binding touch to everything: Learning unified multimodal tactile representations. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024: IEEE; 2024. p. 26330-26343.
39. Li, S.*, et al.* Biomimetic multimodal tactile sensing enables human-like robotic perception. *Nat. Sens.* **1**, 52-62 (2026).
40. Sun, H.*, et al.* A spike–language dual framework bridges fast perception and deep reasoning in artificial tactile somatosensory systems. *Nat. Sens.*, 1-12 (2026).
41. Xu, Z.*, et al.* UniT: Data efficient tactile representation with generalization to unseen objects. *IEEE Robot. Autom. Let.* **10**, 5481-5488 (2025).
42. Wang, J.*, et al.* TempTrans-MIL: A general approach to enhancing multimodal tactile-driven robotic manipulation classification tasks. *IEEE/ASME Trans. Mechatron.* **30**, 3915-3926 (2025).
43. Johannsmeier, L., Schneider, S., Li, Y., Burdet, E., Haddadin, S. A process-centric manipulation taxonomy for the organization, classification and synthesis of tactile robot skills. *Nat. Mach. Intell.* **7**, 916-927 (2025).
44. Yang, R.*, et al.* Iontronic pressure sensor with high sensitivity over ultra-broad linear range enabled by laser-induced gradient micro-pyramids. *Nat. Commun.* **14**, 2907 (2023).
45. Lee, J.-H.*, et al.* Hypersensitive meta-crack strain sensor for real-time biomedical monitoring. *Sci. Adv.* **10**, eads9258 (2024).
46. Nakamura, N., Iwase, E. Stretch-based kirigami structure with folding lines for stretchable electronics. *Npj Flex. Electron.* **9**, 51 (2025).
47. Li, Y.*, et al.* Multifunctional biomimetic tactile system via a stick-slip sensing strategy for human–machine interactions. *Npj Flex. Electron.* **6**, 46 (2022).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.docx